

Digital Methods for Internet Research

Part II

4. Website histories and historiographies

N. Brügger (2009). "Website history and the Website as an object of study." *New Media & Society*. 11(1-2): 115-132.

B.A. Howell (2006). "Proving web history: How to use the Internet archive." *Journal of Internet Law*. 9(8): 3-9.

J. Murphy, N.H. Hashim & P. O'Connor (2007). "Take me back: Validating the Wayback Machine." *Journal of Computer-Mediated Communication*. 13(1).

S. Schneider and K. Foot (2004). "The web as an object of study." *New Media and Society*. 6(1): 114-122.

M. Thelwall, L. Vaughan and L. Björneborn (2005). "Webometrics," in B. Cronin (ed.), *Annual Review of Information Science and Technology*, 39, Medford, NJ: Information today, 81-135.

5. Cross-spherical analysis and other forms of comparative media studies

R. Rogers and N. Marres (2002). "French scandals on the Web and on the street: Experiment in stretching the limits of reported reality." *Asian Journal of Social Science*. 30(2): 339-353.

Richard W. Wiggins (2001). "The Effects of September 11 on the Leading Search Engine." *First Monday*, 7(10),
<http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/890/799> (accessed 18 October 2010).

D. Weaver and B. Bimber (2008). "Finding news stories: A comparison of searches using LexisNexis and Google News," *Journalism & Mass Communication Quarterly*. 85(3), 517-532.

B. Etling, J. Kelly, R. Faris and J. Palfrey (2009). "Mapping the Arabic Blogosphere: Politics, Culture and Dissent." Berkman Center Research Publication No. 2009-06, Cambridge, MA: Harvard University.

6. Wikipedia & networked content analysis

M. Burke and R. Kraut (2008). "Taking up the mop: Identifying future Wikipedia administrators," *Proceedings of the 2008 CHI Conference*, April 5-10, Florence, Italy, 3441-3446.

T. Chesney (2006). "An empirical examination of Wikipedia's credibility," *First Monday*, 11(11), November.

S. Niederer and J. van Dijck (2009). "The case of Wikipedia: Wisdom of the crowd or technicity of content?" unpublished ms.

J.M. Reagle, (2008). In good faith: Wikipedia and the pursuit of the universal encyclopedia, PhD dissertation, New York University, pp. 70-102.

4. Website histories and historiographies

N. Brügger (2009). "Website history and the Website as an object of study." *New Media & Society*. 11(1-2): 115-132.



Website history and the website as an object of study

NIELS BRÜGGER

University of Aarhus, Denmark

Abstract

This article puts on the agenda one of the fundamental theoretical questions within the emerging field of website history: how can the object of historical study – the website – be delimited? Its focus is on the ‘website’ artefact as a medium and a text. After elaborating a definition of the website, as well as discussing how the website is distinct from other possible analytical web objects (the web as a whole, web sphere, webpage and textual web element), the article addresses the challenges of delimiting the archived website. Finally, it outlines some of the key issues in a general discussion of website history.

Key words

medium • text • web • web archiving • web history • web sphere • website

Today, no one would dispute that the internet has been an important part of our communicative infrastructure for some years. Nevertheless, internet history is a relatively blank page, not to mention the subdiscipline of website history, which can be considered an emerging discipline at the intersection between media history and internet history, and which regards the individual website as the unifying entity of historical analysis. Although the number of website studies of all kinds has been growing for several years, genuine historical studies of websites are almost non-existent.¹

This article puts on the agenda one of the fundamental theoretical challenges within the field of website history: how can the object of historical study – the website – be delimited, both on the web and as an archived website? The aim is to contribute to the development of a theoretical framework enabling the website historian to discuss if and how the bits and pieces of web material that can constitute a website are kept together, so that they form a coherent and delimited object of study.²

GENERAL APPROACH

In order to be more specific about the kind of theoretical framework to be discussed, it may be fruitful to introduce a distinction between two types of theories, depending on what they aim at understanding.

On the one hand, we have what could be called theories of objects, which are directed towards how the object of study can be delimited in a systematic and consistent way: for example, theories of what can be understood by ‘media’, ‘communication’ and ‘the social’. On the other hand, we have theories of explanation, which try to explain and understand the object: for example, the use of semiotics, rhetorical theories, discourse theories, theories of social interaction and the like, which are used with a view to explaining the function and use of media, communication or the social. It can be argued that the former type of theory precedes the latter: we have to know what we study before we can explain it. Thus, since this article sets out to reflect on how to identify the website as an analytical object, it can be regarded as an outline of a theory of the object ‘website’, while other kinds of theoretical approaches must follow in order to explain the various aspects of websites in more detail. Therefore, the following theoretical framework as well as the reflections on the archived website might be relevant also for website analysis in general, not only for website history. First, website studies need to be more precise about their object of study; second, in practice, most website studies are based on archived websites in order to have a stable object to study and to wcto refer.³

Possible analytical objects of media studies

If website history is considered to be a natural part of media history, in continuation of the history of newspapers, film or television, for example, an obvious starting point for the theoretical discussion is the possible delimitations of the object of study within media studies and media history in general.⁴

When faced with a given media artefact, media studies can focus on five different constituents: sender, medium, text, receiver and context, and the mutual relations between these. Although these constituents resemble a communication model, they should be understood as a schematic

representation of the possible analytical objects of media studies and not as a model of communication, seeking to explain the elements and phases in a communicative process (see Brügger, 2002b). For example, with regard to 'television', the media scholar can decide to study the media organization (sender), the technical devices used for production, distribution and reception (medium), the reception of programmes (receiver), specific programmes or the flow of programmes (text), the societal context (macro-context) or the immediate context of use (micro-context). Finally, the relations between the above-mentioned constituents can be studied, for example, the question of how the technical side of the medium affects the producers, the text and its reception.

In particular, the medium itself is important, partly because it binds all the other objects together and partly because its material manner of existence sets up the frameworks for its use in offering a media-specific field of possibilities, thus making possible (or impossible) different ways of acting as sender or receiver or of textuality 'within' the medium.⁵ For example, live TV always has been a possibility in making television, and in the early years this potential was used whenever what was aired was not kept on film, since videotapes had not yet been invented. With the advent of tapes the medium changed and live TV began to play a less predominant role, although it remained a potential for many years. This potential was actualized again and used extensively with the invention of satellite television (of course, many economic, organizational and other factors played a role in this development).

Taking these general considerations as a point of departure, it becomes clear that the question of how to delimit the website as an object of study is related primarily to the manner of existence of the 'website' as an artefact, which is why the present focus must be on the medium and the text rather than on the sender, receiver or context. One might argue that leaving the users (sender or receiver) out of consideration is problematic when dealing with the website, since the existence of hyperlinks implies that one has to take the concrete movements of the user into consideration in order to understand the website. This article would argue, first, that this question is not specific to the web, and second, that the signifying elements in newspapers, books, television or on websites can be studied without involving a concrete reader or viewer's way of turning the pages, looking at it or jumping from one window to another. What we study, then, is the text as it presents itself as a field of possibilities to be used (a study that should be an integrated part of any study of the actual use). The user always confronts a text that is already structured in a specific way with specific possibilities, no matter how it is actually used by the user. Thus, the concrete use of the text is considered to be a possibility for an analysis of the medium and the text, rather than a necessity. It could be maintained that when the website historian studies the

text as a field of possibilities, they actually take the viewpoint of the site visitor. However, the website historian's intention is not to visit the website or to re-establish the visitor's experience; rather to analyse the website, at first with a view to delimiting it as an object of study. Therefore, at this stage the website is not approached as an entity experienced by a concrete visitor, but as an object of analysis in its own right. The result of this can be used later as a stepping stone to analysing how the website was actually experienced, often based on other sources than web material.

Medium and text

However, when the website as a medium and a text is in focus, one has to reflect on how it is possible to draw a clear line between the two. This article would argue that the distinction lies in the approach taken by the media scholar, rather than it being inherent in the media artefact. Faced with a newspaper, for example, the focus can be on the paper, the binding and the ink as either physical, material objects, or as an object that through lines, shapes and colours form units endowed with significance (e.g. characters, still images); finally, the focus can be on the relation between material object and signifying units insofar as the paper and ink enhance the use of letters and exclude the use of moving pictures and sound (see Brügger, 2002b). The artefact is the same, but in the first case we consider it a medium, and in the second a text, the difference being whether it is being focused upon as signifying or not.

However, this clear distinction is complicated by digital media, since with the computer a part of the medium is also a text. Today the computer uses an alphabet with two letters (0 and 1) which can be combined on different levels in accordance with certain syntactic rules, thus constituting the grounds for what we see on the screen or hear via the speakers (see Finnemann, 1999). Therefore, an analysis of the computer can focus on the same artefact in three ways. First, the computer-as-medium, that is, the 'box' with the computer, the screen and dots of light and colours on it, the speakers and so forth; second, the computer-as-text, or the dots of light and colours on the screen or soundwaves in the speakers as organized elements of significance; third, the computer-as-medium/text, or the intermediate textual level that the codes of 0/1 and their syntax constitute – a level that is in itself a text, insofar as it is composed of letters and a syntax (see the discussion of materiality and immateriality in Brügger, 2002a).

In this way, digital media are distinctive by embodying a level which can be viewed as both medium and text. However, the text written with the letters 0/1 is not visible and decipherable to the end-user; this is only the case with the text on the screen that is its product. Further, if the distinction between medium and text is based on the fact that in order for something to

be a text it has to be signifying to a reader, then the text written with the letters 0/1 is part of the medium rather than of the text. Still, this by no means implies that this medium/text is unimportant; on the contrary, it is highly important, since as part of the medium it is a decisive part of the conditions for the textual elements.

ELEMENTS OF A WEBSITE ANALYSIS

In light of these general considerations it is argued that, at minimum, the following three elements should be part of a historical analysis of the website:

- its media environment;
- its textual environment; and
- its textuality, that is, the textual elements that constitute the website as a textual phenomenon *strictu sensu*.⁶

Media environment of the website

The media environment in which the website is embedded is the internet, and it must be included in historical analysis of the website since it constitutes the array of possibilities and constraints that frames the other two elements, i.e. the textual environment and the textual elements that constitute the actual website.

Although the internet might seem of no particular interest for a study of the website, its specific manner of existence becomes immediately apparent when undertaking historical studies of five or 10-year-old websites. In 1995, graphics were not in general use, just as streaming of sound and moving images were not possible, and in order to be able to explain this, the internet must be implicated. This article proposes an analysis of the internet as a medium by tracing the historically specific articulations of what can be considered the invariant traits of the internet.

All forms of the internet have the following three characteristics. First, computers must be a component; second, computers are directly connected; and third, when connected, they are able to recognize one another and communicate by means of a shared system of addresses and a shared language: a protocol. Since computers are an essential part of the internet, it 'inherits' the invariant properties characterizing the computer: a mechanical (binary) alphabet consisting of a finite set of letters, each of which is void of semantic content; an algorithmic syntax; and an interface (cf. Finnemann, 1999). None of these invariant traits are necessary in a specific form, but they are necessary as such.

These invariant traits will always be materialized in concrete artefacts that are historically, socially and culturally specific. Concerning the computer (the first of the three traits) we must study the mechanical alphabet (for the computers of today there is not much to say on this point; the alphabet, which

is binary: 0/1); the algorithmic syntax, which varies with the development of software (applications and programs on different layers); and the interface, where some of the following points are of interest: the 'box' with the computer, the screen, the pointing or clicking device, the keyboard, visual in, audio in, the sound source and various storage devices. Concerning direct connectedness, the following are of interest: the devices that connect, the computers that act as 'intermediate stations' by carrying out tasks in relation to the direct connectedness, of which some are specific to the web, and the actual architecture of the connectedness. Finally, concerning the shared system of addresses and shared language, one could focus on internet protocols, protocols for transfer, identification and localization and various shared languages.

Textual environment of the website

The website is not only embedded in a media environment, but also in an environment of signifying elements, or rather two textual environments. The first of these is related to the use of the computer as such, namely, the graphical user interface, with its two key elements which are of great importance for understanding the website: the desktop with clickable and movable icons, folders and the like which serve as a background for the website, and scalable, movable and overlapping windows.

In addition to the desktop and window, which are there no matter what the computer is used for (at least in the case of most personal computers for some years now), the website is embedded in a textual environment that is a necessary and specific condition for the use of the web, namely, the web browser and additional software (however, both can be used for browsing and opening files on the computer). The web browser makes it possible to communicate with web servers, using transfer and location protocols (e.g. http, URL) and to translate what is written in a shared language (e.g. HTML) to letters, images, sounds and positioning in the window. The web browser in itself normally features textual elements that might be of importance in order to analyse a website: for example, the 'Back' button, the 'Go to/History' in the menu bar and the address bar with the URL that all can be elements of navigation *de facto*, even if as such they are not part of a specific website. The variety of additional software allows for the processing of different elements of expression (e.g. QuickTime player, Windows Media players, Adobe Acrobat Reader), and these all constitute textual environments of their own (words, graphics) which are not part of a specific website.

As was the case with the media environment, the textual environment might seem of little interest, but it nevertheless plays a role as a framing facility, which again historical studies will surely make us recognize, since the desktop, browsers and additional software have changed significantly in the course of history.

Textuality of the website

When confronted with a fragmented variety of textual elements that come and go in our browser windows, how can we draw clear demarcation lines that make it possible to delimit the individual website? In order to identify the textual constituents of the web as well as their coherence, this article discusses the following basic concepts: textual element; morphology and syntax; and levels of morphological and syntactical interrelations.

The basic unit of analysis is the textual element, which is a defined coherent textual unit composed of one of the following four formats of expression: written letters (or other characters), still images, moving images and sounds. Textual elements could be a headline and the body text (writing), a photograph (image), a banner advert, a news story from television (moving images), or a piece of music (sound). A textual element can be composed of smaller textual elements (for example, body text can consist of several paragraphs, which can be considered to be individual textual elements).

Analysis of the textual elements can unfold in two dimensions: morphological and syntactical. The morphological analysis focuses on the characteristics of the individual element: how and by what means is a still image constructed? How are the interior structure and coherence of written textual elements established? The syntactical analysis focuses on the rules governing the combinations of elements, as well as on the functions of these combinations, in the sense of the relations between identical types of elements as well as varying combinations (the relation between headlines and body text, images and captions, or written words in images, etc.); and finally, the syntactic dimension can focus on the overall combination of all the types of elements in the complete audiovisual composition.⁷

The interrelations between several textual elements (syntax) as well as the coherence within an individual textual element (morphology) can be studied on three different levels. First, the relations between textual elements (or the limits of an individual textual element) are established through semantic interrelations: for example, by the use of textual cohesion and coherence; the grammatical, lexical construction of a coherent 'world', isomorphy between semantic entities and a global menu structure. Second, the relations between textual elements (or the limits of an individual textual element) are established through formal interrelations, i.e. the forms of expression. For example, the relations between elements of writing are made formally coherent through the formal traits of the words and letters, their positioning (imposed lines of the same length), typography (font, colour, size) and the immediate environment (different forms of separators such as vertical or horizontal lines, boxes, scroll bars); or, for example, the limits of the image element are established by different types of framing. Third, the relations between textual elements (or the limits of an individual textual element) are

established through physically performative interrelations, i.e. the possibility of physically performing an action or a movement, often between elements that are not immediately visible. This could be different forms of continuous movements (vertical or horizontal scroll) or discontinuous 'jumping' (click, hyperlink, mouse-over); thus, the physically performative relationship adds a concrete possibility of action to the other textual elements. However, it is always dependent on the ability to 'graft' onto something which has been expressed already semantically or formally – a word, an image – which is why, for example, a hyperlink never exists without a textual 'substantiation' and must be understood in close relation to this. So moving around in or between textual elements is a semantic, formal as well as a physically performative activity.⁸

With this vocabulary in mind, it can be maintained that from a textual point of view, the website is a coherent textual unit that unfolds in one or more interrelated browser windows, the coherence of which is based on semantic, formal and physically performative interrelations (the same criteria can be used to draw internal demarcation lines in a website, thus delimiting subsites). It could be added that the more widespread these three types of interrelations are, the greater the coherence of the website. Thus, when the website historian aims to characterize the variety of web material in their browser windows, they should try to answer the following three questions: does the web material treat the same subject (semantic cohesion)? Does it look alike (formal cohesion)? Is it possible to come from one browser window to another (physically performative coherence)? If these three questions are answered in the affirmative, we can call the web material a website. However, one of the answers might be a 'Yes, but ...', and in this case we might still consider the web material a website if, for example, it consists of a 'core' of coherent webpages with some loosely attached webpages 'around' it, or it is not very coherent but coherent enough to be considered a website. In this sense, coherence is a matter of degree.

THE WEBSITE AND STRATA OF THE WEB

The above theoretical framework can assist the website historian in the task of discussing and maybe even determining to what degree the confused mass of content in web browser windows form a website. It also makes it possible to delimit the website from other possible analytical entities on the web. Within the variety of web material that constitutes 'the web', one can distinguish the following five analytical strata: the web as a whole; the web sphere; the individual website; the individual webpage; and an individual textual web element on a webpage, such as an image.

The concept 'web sphere' was coined by Schneider and Foot and is defined as 'not simply a collection of websites, but as a set of dynamically defined

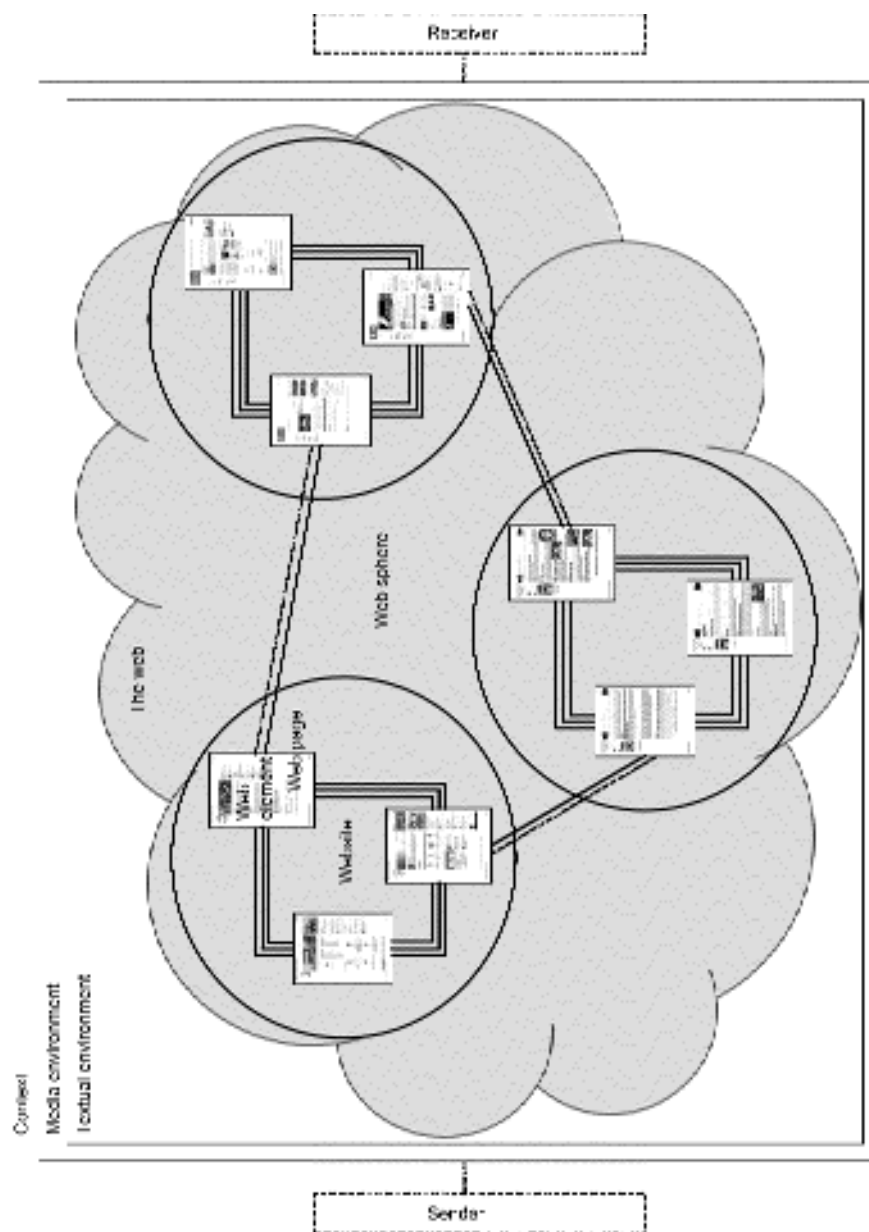
digital resources spanning multiple websites deemed relevant or related to a central event, concept, or theme' (Schneider and Foot, 2006: 20). For example, a web sphere could be web activity in relation to a political event such as the US elections of 2000, 2002 and 2004 (see Schneider and Foot, 2006). As the quotation shows, the web sphere is constructed by the researcher, by selecting websites based on their relevance to the theme in question. On a general level, this is in accordance with the use of the concept in the present article. However, in contrast to Schneider and Foot, the main concern of the present theoretical framework is not the use of websites by actors, but how the researcher can determine that the concrete web material before them is a website, and that it treats the theme in question. The starting point of web sphere analysis – that is, the identification of the web objects of the web sphere (see Schneider and Foot, 2006) – must be based on systematic reflections on how the website as well as the themes are established and supported by textual means such as coherence and creation of meaning (semantics).

Among the five clusters of textuality mentioned above, the website stands apart in the following ways (see Figure 1).

It is clearly distinct from the web as a whole, since the web as such is not supported by any kind of consistent semantic, formal or physically performative interrelations. The website is also clearly distinct from the web element and the webpage: the web element is the minimal textual element on a webpage, while the webpage is clearly delimited by the individual browser window (rarely are webpage and website identical; however, a website consisting of only one webpage is a possibility). Finally, the website is distinct from the web sphere insofar as the website is based on a semantic, formal and physically performative coherence, while the web sphere is characterized only by partial semantic and/or performative interrelations (see the two lines between websites in Figure 1). The 'digital resources' that span 'multiple websites' (Schneider and Foot, 2006) constitute a coherent analytical object because they share a semantic world (the 'central event, concept, or theme' that they treat) and/or because they are linked together; but the relations between them are weaker than those of the website, partly because there are no formal relations and partly because neither the semantic nor the performative relations are necessarily expressed in the same explicit manner on all of the websites that are part of the web sphere.

In summary, the website can be delimited as an analytical object by stating that it is a coherent textual unit which:

- unfolds in one or more interrelated browser windows, the coherence of which is based on semantic, formal and physically performative interrelations;
- is established on the background of, and in an interplay with, a certain internet and textual environment; and



• Figure 1 The elements of a website analysis

- is clearly distinct from the web as a whole, as well as from the web sphere, webpage and web element.

It could be argued that the website can be delimited by other means, such as the web server, the producer of the website or the domain name. However,

the website is not always consistent with the web server and vice versa, since one web server can host several websites, just as one website can span several web servers. The producer of the website can be seen as a unifying editorial entity that exerts overall editorial control and coordination over the webpages, but this editorial role is only useful as a demarcating mechanism insofar as it is expressed on the website through, for example, the use of a logo, footer or disclaimer on each webpage, or by means of a consistent link structure: in other words, insofar as it is expressed by textual means. In fact, the same goes for the domain name: it can constitute a unifying element, but it is a textual web element in line with the other textual elements that create the unity of a website. Although the web server, producer of the website and domain name cannot be considered to be the demarcating mechanism of the website, all three constitute important supplements to the cohesive agent that is created by the semantic, formal and physically performative relations between textual elements.

THE ARCHIVED WEBSITE

Even if the theoretical vocabulary outlined above can help the website historian in the task of delimiting the website as an analytical object in general, this step forward is challenged immediately when the object of historical study is accessible only as an archived website, which is almost always the case. As will be shown later, the archived website differs significantly from the website on the live web as well as from other known archived media types (by 'live web' it is meant 'what is on the web today'). Therefore, the question 'How can the website be delimited as an object of study?' must be supplemented by the question 'How can the archived website be delimited?'⁹

No matter how an archived website has been created and no matter which archive it is found in, the website historian cannot expect it to be an identical copy on the scale of 1:1 of what was actually on the live web at a given time. This is because, on the one hand, the archived website is an actively created and subjective reconstruction, and on the other hand, it is almost always deficient. The combination of these two characteristics has a bearing on the extent to which it is possible to delimit temporally the archived website in a consistent manner.

An actively created and subjective reconstruction

An archived website is always an actively created subjective reconstruction. First, the simple fact that a choice has to be made between different archiving forms and strategies, both in general and in detail, implies that the archived website is based on a subjective decision of either an individual or an institution (see Brügger, 2009b, forthcoming, on different archiving forms and

strategies). For example, it must be decided which form of archiving is to be used, where the archiving should start, how far away from the starting point the archiving is to continue (both in depth and breadth), whether specific file types are to be included or excluded (e.g. images, sounds, flash), whether material is to be collected from other servers and how the material is to be preserved, both here and now and in the long-term perspective. Second, the archived web document is a reconstruction in the sense that it is recreated on the basis of a variety of web elements that stem from the live web and which are reassembled and recombined in the archive.

Thus, the archived website is the result of an active process taking place at the nexus of the 'raw material' which is present on the web and a number of choices with regard to selecting and recombining the bits and pieces at hand. In this sense the archived website does not exist prior to the act of archiving; it is only created in a stable form through the archiving process. Thus, the archived website stands apart from other media types. When archiving newspapers, film, radio and television, the main choices are related to selection of the material, while the archiving process itself almost exclusively consists of taking a copy out of circulation and storing it; no matter who stacks the newspapers or presses the record button on the video recorder, the archived copies are identical to what was once in circulation, just as all copies are identical. In contrast, with websites, choices have to be made in relation to selecting and archiving and we always do more than just remove the web material from circulation; the material is never totally unchanged.

Deficiencies

When compared to what was on the live web, the archived website is almost always deficient. Apart from the deficiency caused by deliberately chosen omissions, two other sources of deficiency can be singled out: those caused by technological problems during the process of archiving, and those related to time. The archived website is very likely to be deficient for technical reasons (software or hardware). For example, words, images or graphics, sounds and moving images can be missing, or some of the possibilities of interaction can be inoperative in the archived web document.¹⁰

Apart from the technical problems, one of the major reasons for deficiency is related to what could be called the dynamic of updating, i.e. the fact that the web content might have changed during the process of archiving and we do not know if, where and when this happens (see Brügger, 2005; Masanès, 2006; Schneider and Foot, 2004). A brief example can illustrate this:

During the Olympics in Sydney in 2000, I wanted to save the website of the Danish newspaper *JyllandsPosten*. I began at the first level, the front page, on

which I could read that the Danish badminton player Camilla Martin would play in the finals a half hour later. My computer took about an hour to save this first level, after which time I wanted to download the second level, 'Olympics 2000'. But on the front page of this section, I could already read the result of the badminton finals (she lost). The website was – as a whole – not the same as when I had started; it had changed in the time it took to archive it and I could now read the result on the front page, where the match was previously only announced. (Brügger, 2005: 22–3).

As this example illustrates, it is obvious that the archived website is deficient, since it is incomplete compared to what was once on the live web: something is lost in the process of archiving due to the asynchrony between updating and archiving. However, it is also deficient in another and less obvious way, in that the archived web document is not only incomplete, but is also 'too complete': something that was not on the live web at the same time, the content of two webpages or website sections, is now combined in the archive and it is difficult to determine what the website was actually like at a given point in time on the basis of these two entities.

In summary, it can be argued that since the actual act of preserving web material almost always changes the material that was on the web in a number of ways, the process of archiving creates a unique version, but not a copy.¹¹

Temporal delimitation of the archived website version

As a consequence of these characteristics, a general question presents itself: how can we temporally delimit an archived version of a website? In fact, on the live web, the question of whether a website is 'the same' version or not in temporal terms is not relevant. If we open a webpage and jump to another webpage on the same website by means of a hyperlink, the two webpages are part of the same version simply because they are there at the same time. However, this is not the case with an archived website, since the two webpages might have been archived at different points in time and several archived copies of each of them might even exist, archived at, for example, five-minute intervals. Instead of being strictly related to a distinct point in time, the archived version extends like a time continuum, thus making it difficult to determine not only a temporally coherent version, but also the limits between versions, since in fact one and the same web element can be part of more than one version. The archived websites might appear to be continuous versions with blurred and overlapping borders.

The website historian, then, faces the task of determining whether a variety of webpages are part of one and the same version or not, thus questioning the temporal consistency as well as the temporal limits of the version(s). In this

task they are obliged to choose between two paths. Either a very strict criterion is maintained, that there should be simultaneity between all the archived web elements, or a more pragmatic criterion is accepted, according to which some kind of interval between the web elements is tolerated. In other words, either the version is delimited very precisely, with the result that a lot of versions will consist of only one webpage, or the version is delimited in a more imprecise and inconsistent manner, thereby creating fewer and more 'complete' versions.

To conclude: on the live web the main problem is to delimit the website in spatial terms, while the temporal limits of the website are simply present. However, things work differently when a website archive is entered. Here, not only the spatial but also the temporal limits must be established by the researcher and to a certain extent the archived version is constituted and constructed only *post festum*, after the archiving process. In other words, the subjectivity, active reconstruction and temporal dynamics of updating which characterize the moment of archiving are mirrored in the archive, thus forcing the website historian to delimit the archived website as well as to formulate in detail the methodological deliberations that enable them to do so.

CONCLUSION

The theoretical framework and reflections on the archived website formulated above can be considered a conceptual toolbox that enables the website historian – and the website scholar in general – to discuss to what degree the web material on their screen forms a website in order to create a consistent object of study, as well as to discuss the borders of a website over time – that is, if it had the same boundaries two, five and 10 years ago. However, it can serve also as a starting point for more general reflections on website history. These concluding remarks will outline some of the key issues in this general discussion.

One of the tasks that could be undertaken would be to formulate a historical textual grammar of the website. When the website is defined as above, it would be relevant to ask how the textual elements that constitute the website, along with the semantic, formal and physically performative relations between them, actually have appeared and functioned at various periods in the past in order to identify recurrent patterns and traits. In other words, it is important to examine the ways in which the textual web constituents as well as their combinations actually have been used, so as to formulate the historically changing morphological and syntactical rules of website textuality.

Furthermore, one could point to a cluster of tasks that revolve around a reopening of the different themes which, until now, have been incidental to the focus on specific characteristics of the website. These issues can be

addressed now with a systematic approach to the specificity of the website in mind.

First, it should be discussed whether well-known theories of explanation can be applied unchanged to an object of study such as the website. What does it mean to do rhetorical studies or discourse analysis when, for example, the relations between the fragments of content and expression are performed physically by a hyperlink from one browser window to another, rather than by turning pages or looking at the changing pictures on a television screen? Or when an image is used as a point of departure for a hyperlink, thus making it possible for an image to refer to other textual elements all by itself and in a physically performative sense? Or when written words or pictures suddenly appear with the use of a mouse-over?

Second, both the actual users (sender and receiver) and the context of the website should be reintegrated into historical study, with the historically-specific mediacy and textuality of the website in mind. What role have mediacy and textuality played as fields of potentialities for website users at various points in time in the past? How have the users actually interacted with these possibilities? How has the context of the website – other media, for example – interacted with the website (for example, various cross-media phenomena)?

Third, the website should be resituated in its relation to the other strata of the web, i.e. the web as a whole, web sphere, webpage and web element. These five strata are one another's mutual context, and in many ways they are interdependent, but it could be discussed whether one or more of them constitutes the analytical centre of gravity in historical study of the web. On the basis of the theoretical reflections above, this article would suggest that the website and web sphere be considered the two overriding analytical units on the web, since the other strata are subordinate to them to various degrees, and they are both based on a set of semantic and physically performative relations. The individual web element is difficult to explain without the webpage as its context, just as the webpage is almost always part of a website. The web as such is nothing but a fragmented, heterogeneous and confused mass, if one does not draw some semantic, formal or physically performative lines in it (see Brügger, 2007). However, the website and web sphere are not on equal terms. One may argue that it is possible to analyse an individual website without relating it to a web sphere, but it is less likely that the web sphere can be studied without preliminary identification and study of the websites that constitute the web sphere. In this sense, the website tends to be the encompassing analytical unity on the web and within web history.

As these concluding remarks indicate, the field of website history is still in its infancy and more theoretical research is needed in order to establish the research field. The reflections above are but a small contribution to this task.

Notes

- 1 Major contributions to internet and web history are Abbate (2000), Gillies and Cailliau (2000), Hauben and Hauben (1997), Henderson (2002), Naughton (2002) and Poole (2005). Some of the few historical studies of websites are Burns (2008), Martin (2004), Schneider and Foot (2006), Tkach-Kawasaki (2003) and Voerman and de Graaf (1998). In addition, currently the author of this article is conducting a research project entitled 'The History of dr.dk, 1996–2006', where the history of the dr.dk website is to be written.
- 2 In the present context, 'website' is understood as an entity based on the use of protocols and mark-up languages, which originates from the world wide web in the broadest sense rather than other 'clusters' of files on the internet (for example, 'gopher' sites), although they can be analysed probably along the lines outlined in the following.
- 3 Although one could expect that the fundamental question of what should be understood by 'website' was addressed in the ever-growing number of studies of websites or in general studies of the web, this is not the case. On the contrary, the question is rarely addressed systematically and in depth, if at all (see Burnett and Marshall, 2003; Gauntlett and Horsley, 2004; Kaye and Medoff, 2001; Mitra and Cohen, 1999; Schneider and Foot, 2004 and the brief comments on each of these texts in Brügger, 2007).
- 4 The following considerations on the analytical object of media studies take as their starting point the reflections in Brügger (2002b).
- 5 I use the term 'mediacy' for this 'media-ness', the specific being of a medium (see Brügger, 2002b). Mediacy is considered to be a dialectic movement between the potentialities of the medium and the actual use of it. So the medium is both the field of possibilities and the result of its use (see Brügger, 2002b).
- 6 'Textual' means that the level of analysis is 'higher' than letters or phonemes and sentences. The term 'textual' is understood in a broad sense and is not limited merely to written language, rather it refers to all forms of expression such as written words, still images, moving images and sound.
- 7 The terms 'morphology' and 'syntax' are used in a broad sense, extending the usual meaning, where focus is on similar elements (written or spoken words) unfolding in one dimension only (a progressive chain of expression). In this context, all kinds of textual elements as well as relations in several directions are comprised in the words 'morphology' and 'syntax'.
- 8 The word 'performative' is inspired by speech act theories, where the use of a word is the performing of an action. However, in the present context, what is performed is not just a 'speech act' but a physical act: on the web we actually do things with words and images.
- 9 Web archiving means any form of deliberate and purposive preserving of web material (this definition is elaborated in more detail in Brügger, 2009b, forthcoming; for introductions to and discussions of web archiving, see Brown, 2006, Brügger, 2005, 2008, 2009b, forthcoming; Masanès, 2006).
- 10 For the test of the quality of archived web documents in various archives, see Brügger (2009a, forthcoming); see also the discussion of quality and completeness in Masanès (2006).
- 11 This uniqueness of archived versions is confirmed by the test presented in Brügger (2009a, forthcoming), where the appearance of the same website which has been archived on the same date in different archives was examined. Thus, the archived web document must be treated in new ways, compared to other archived media types,

based on a new kind of philology; a small contribution to such a 'website philology' is outlined in Brügger (2008, 2009b, forthcoming).

References

- Abbate, J. (2000) *Inventing the Internet*. Cambridge, MA: MIT Press.
- Brown, A. (2006) *Archiving Websites: a Practical Guide for Information Management Professionals*. London: Facet Publishing.
- Brügger, N. (2002a) 'Does the Materiality of the Internet Matter?', in N. Brügger and H. Bødker (eds) *The Internet and Society? Questioning Answers and Answering Questions*, pp. 13–22. Aarhus: Centre for Internet Research.
- Brügger, N. (2002b) 'Theoretical Reflections on Media and Media History', in N. Brügger and S. Kolstrup (eds) *Media History: Theories, Methods, Analysis*, pp. 33–66. Aarhus: Aarhus University Press.
- Brügger, N. (2005) *Archiving Websites: General Considerations and Strategies*. Aarhus: Centre for Internet Research.
- Brügger, N. (2007) 'The Website as Unit of Analysis? Bolter and Manovich Revisited', in A. Fetveit and G.B. Stald (eds) *Digital Aesthetics and Communication*, pp. 75–88. Bristol: Intellect.
- Brügger, N. (2008) 'The Archived Website and Website Philology: a New Type of Historical Document?', *Nordicom Review* 29(2): 151–71.
- Brügger, N. (forthcoming, 2009a) *Archived Websites between Copies and Versions: Test of Versions in Existing Web Archives*. Aarhus: Centre for Internet Research.
- Brügger, N. (forthcoming, 2009b) 'Web Archiving: between Past, Present and Future', in R. Burnett, M. Consalvo and C. Ess (eds) *Blackwell Companion to Internet Studies*. Oxford: Blackwell.
- Burnett, R. and P.D. Marshall (2003) *Web Theory: An Introduction*. London: Routledge.
- Burns, M. (2008) 'Early ABC Online Interactive Sites. Remembering Public Service Broadcasting: Liberty and Security in Early ABC Online Interactive Sites', *Television and New Media* 9(5): 392–406.
- Finnemann, N.O. (1999) 'Modernity Modernised: The Cultural Impact of Computerisation', in P.A. Mayer (ed.) *Computer, Media and Communication*, pp. 141–59. Oxford: Oxford University Press.
- Gauntlett, D. and R. Horsley (eds) (2004) *web.Studies*. London: Arnold.
- Gillies, J. and R. Cailliau (2000) *How the Web Was Born: The Story of the World Wide Web*. Oxford: Oxford University Press.
- Hauben, M. and R. Hauben (1997) *Netizens: On the History and Impact of Usenet and the Internet*. Washington: IEEE Computer Society.
- Henderson, H. (2002) *Pioneers of the Internet*. San Diego, CA: Lucent Books.
- Kaye, B.K. and N.J. Medoff (2001) *The World Wide Web: A Mass Communication Perspective*. New York: McGraw-Hill.
- Martin, F. (2004) 'Net Worth: The Unlikely Rise of ABC Online', in G. Goggin (ed.) *Virtual Nation: The Internet in Australia*, pp. 193–208. Sydney: University of New South Wales Press.
- Masanès, J. (ed.) (2006) *Web Archiving*. Berlin: Springer.
- Mitra, A. and E. Cohen (1999) 'Analyzing the Web: Directions and Challenges', in S. Jones (ed.) *Doing Internet Research*, pp. 179–202. London: Sage.
- Naughton, J. (2002) *A Brief History of the Future: The Origins of the Internet*. London: Phoenix.

- Poole, H.W. (ed.) (2005) *The Internet: A Historical Encyclopedia*. Santa Barbara, CA: ABC/Clio.
- Schneider, S.M. and K.A. Foot (2004) 'The Web as an Object of Study', *New Media & Society* 6(1): 114–22.
- Schneider, S.M. and K.A. Foot (2006) *Web Campaigning*. Cambridge, MA: MIT Press.
- Tkach-Kawasaki, L.M. (2003) 'Clicking for Votes: Assessing Japanese Political Campaigns on the Web', in K.C. Ho, R. Kluver and K.C.C. Yang (eds) *Asia.com: Asia Encounters the Internet*, pp. 159–74. London: Routledge Curzon.
- Voerman, G. and J.D. de Graaf (1998) 'De websites van de nederlandse politieke partijen, 1994–1998' ['Dutch Political Party Websites 1994–1998'], in *Documentatiecentrum Nederlandse Politieke Partijen Jaarboek 1997* [*Dutch Political Party Central Documents Yearbook 1997*], pp. 238–69. Groningen: Documentatiecentrum Nederlandse Politieke Partijen.

NIELS BRÜGGER is Associate Professor at the Centre for Internet Research, University of Aarhus, Denmark. His primary research interest is website history, and in 2007 he began writing a history of the Danish Broadcasting Corporation's website entitled *The History of dr.dk, 1996–2006*. He has published a number of articles, monographs and edited books, including *Archiving Websites: General Considerations and Strategies* (Centre for Internet Research, 2005) and *Media History: Theories, Methods, Analysis* (ed. with S. Kolstrup, Aarhus University Press, 2002).

Address: Institute of Information and Media Studies, University of Aarhus, Helsingforsgade 14, 8200 Aarhus N, Denmark. [email: nb@imv.au.dk]

B.A. Howell (2006). "Proving web history: How to use the Internet archive." *Journal of Internet Law*. 9(8): 3-9.

PROVING WEB HISTORY: HOW TO USE THE INTERNET ARCHIVE

By Beryl A. Howell

Showing what content a Web site *previously* contained (as opposed to what is currently on the site) may help answer questions that attorneys confront in a myriad of cases, ranging from copyright and trademark infringement to business torts and defamation. Showing that a particular Web site is currently using copyrighted text or images or protected marks may be all that is needed in a case, but documenting prior versions of the Web site can be critical to establish the scope or extent of the illegal or tortious conduct, the amount of the damages, or the requisite *mens rea*.¹ When the Web site at issue or the offending content on it has been removed or modified, the most effective, if not the only, way to document the content is to review prior versions stored in online archives of Internet sites.

Specifically, in trade secret and misappropriation cases, showing that the same information claimed to be secret or confidential has previously been made publicly available by the claimant, such as on the claimant's Web site, can be probative if not case dispositive. Diligently searching archived versions of the claimant's Web site for such evidence can be worthwhile.

Similarly, capturing evidence from archived Web sites is helpful in intellectual property infringement cases as well. For example, in cybersquatting and typo-squatting cases, where a trademarked name or a slightly misspelled trademarked name (e.g., *mcrosoft.com*) has been registered as a domain name and used as an online address for a Web site, evidence from archived versions of the Web site may establish the period of time the offending Web site has been operational and the types of goods or services

offered on the site over time. This evidence can help establish intent and harm. In cyberstuffing cases, where popular trademarked names are repeatedly embedded in hidden metatags and transparent text on a Web site, search engines will pick up on the trademarked names and push the infringer's Web page to the top of search engine results, diverting business from the trademark owner's site. Even if the Web site is modified after the infringer is notified of the claim, documenting the cyberstuffing activity on archived versions of the Web site can establish the nature of the offending activity, its scope, and duration.

No matter the legal context, gathering evidence of prior versions of Web sites should be performed in a careful forensic manner with cognizance of the underlying technology used in the archiving process. This article will review strategies and methods for capturing prior versions of Web sites from the most popular of the archives and considerations that counsel should be prepared to address in authenticating this evidence.

"MAP" OF ARCHIVES

At the outset, archived versions of Web sites are available for free at multiple sites. The federal government, in particular, archives government Web sites and makes those archives accessible online. For example, the US Government Printing Office, in partnership with the University of North Texas, provides online access to federal Web sites that have ceased operation on a site called the CyberCemetery.² The archived deceased Web sites include Access America, Advisory Commission on Intergovernmental Relations, Office of Technology Assessment, and others. Similarly, the Electronic Research Collection (ERC),³ which is a partnership between the United States Department of State and the Federal Depository Library at the Richard J. Daley Library, University of Illinois at Chicago (UIC), makes available the US Department of State Web pages archived from 1998 through January 2001. In addition, the National Archives and Records Administration harvested all of the federal agency public Web sites as they existed at the end of the presidential term on January 20, 2005, and makes these archives available at the 2004 Presidential Term Web Harvest site.⁴

Archives of Web sites that are not associated with the federal government are available at several sites. The Library of Congress sponsors a project called Minerva (Mapping the INternet Electronic Resources Virtual Archive), which harvests Web sites based on subject matter and then provides the collections as an archive, rather than try to harvest every Web site. The collections currently available include: Election 2002 Web Archive (July

Beryl A. Howell is a Partner and heads the Washington, DC, office of Stroz Friedberg, LLC, a computer forensics and electronic discovery consulting and technical services firm with offices also in New York City, Minneapolis, and Los Angeles. She formerly served as a New York federal prosecutor and General Counsel of the US Senate Committee on the Judiciary. Research assistance for this article was provided by Donald Allison and Jessica Reust, computer forensic examiners, and George McLean, Evidence Technician at Stroz Friedberg, LLC.

1, 2002-Nov 30, 2002);⁵ September 11, 2001 (September 11, 2001-December 1, 2001);⁶ and Election 2000 (August 1, 2000-January 21, 2001).⁷

Certain Internet search engines, such as Google and Yahoo, also make archived Web sites available. The Google archive provides access to the last cached version of a Web site, but not to prior versions. These cached Web sites are a backup in case the original page is unavailable and are useful since they show the date and time stamps for when each page on the site was retrieved by Google. Google and other search engines often index a Web site about once a month, but Google explains that the “cache is the snapshot that we took of the page as we crawled the web” and cautions that “[t]he page may have changed since that time” or “[t]his cached page may reference images which are no longer available.” Google states that many factors affect how often it indexes a site, but a 2003 survey showed that Google revisited most sites within one month.⁸ Therefore, unless a page is defunct, a Google cached site often will be 30 days old or less. To look farther back in time, the Internet Archive is probably a better bet. Sites may not be cached if they have not been indexed or if the owners have requested that the content not be cached. The date-time stamps on the Google archive may be helpful in establishing, for example, when a site stopped operating within the last six months. If a site is no longer available online, a visit to the Google cache may indicate the date when the site was last indexed.

Yahoo has recently added the ability to view cached pages by clicking on a link entitled “cached.” As with Google, clicking on “cached” brings up a copy of the Web page as it appeared when it was last crawled by the search engine. By contrast to the Google cached sites, however, the Yahoo archive does not date-stamp the version of the cached site but simply notes the following: “It’s a snapshot of the page taken as our search engine crawled the Web. The Web site itself may have changed.” To check the previous versions of the Web site, Yahoo directs users to the Internet Archive. As discussed in more detail below, the Internet Archive contains the most extensive archive of Web sites in terms of period covered, number of Web sites and pages archived, and the number of prior versions of Web sites archived.

Other search engines that provide cached Web sites include *search.msn.com* (MSN), *ask.com* and *teoma.com* (both from Ask Jeeves), *clusty.com* (from meta-search engine Vivisimo), and *Gigablast.com*. Of these, Gigablast may be the most helpful in researching historic Web sites because its search engine results include the date that the Web page was last modified, as well as the date that the page was last indexed by Gigablast. Gigablast also provides links to the cached site, a stripped version of the

site without graphics, and a link to “older copies” found on *archive.org*.

THE INTERNET ARCHIVE AND THE WAYBACK MACHINE

The Internet Archive⁹ is a free online resource that was created in 1996 to build a digital library of Web pages and other cultural artifacts in digital form with the purpose of offering permanent and free access to researchers, historians, scholars, and the general public.¹⁰ Internet Archive provides not only an archive of websites but also of open source movies, feature films, cartoons, historic newsreels, and news video and music.

Five years after its creation, in October 2001, the Internet Archive launched the Wayback Machine, which provides the public with a free online service to search for and access archived Web sites. The name of the search service is derived from the Rocky and Bullwinkle cartoon in which the characters of a bow-tied dog, Mr. Peabody, and his boy assistant, Sherman, used a time machine called the WABAC Machine to travel back in time to famous events in history.

The Web pages are collected for the Internet Archive using a search engine technology called Alexa Crawl that traverses the Internet taking snapshots of Web sites. The Alexa Crawl currently captures about 1.6 terabytes (1600 gigabytes) of Web content per day and takes about two months to complete a snapshot of the more than 16 million Web sites accessible online.¹¹ This search-and-copy engine is owned and operated by Alexa Internet, a for-profit company that offers a free toolbar and a number of statistical services to subscribers based upon the Web content and usage information collected. The company donates a copy of each crawl of the Web to the Internet Archive, which may make the crawl results available after six months. Thus, there is a six- to 12-month lag between the date that a site is crawled and when it appears for free use in the archives of the Wayback Machine.¹² Alexa Internet is now offering a fee-based service to access its crawl results data before it goes to the Internet Archive.¹³

The Alexa Crawl does not purport to capture all Web sites accessible on the Internet, but instead prioritizes the Web sites and pages to copy based on the number of times that a Web site is requested through the Alexa search engine. Thus, not every Web site has an equal chance of being copied at all or copied in full. Alexa Internet uses a rating system for content at all that will be captured. Content that is not popular may be deliberately omitted if not visited often. This is related to Alexa’s business model for selling databases of frequently visited sites to customers. The result is that the Wayback Machine does not hold

archived versions of all Web sites of copies of every page for the Web sites that are archived.

In addition, sites may not be archived if they are password-protected, the site owners have requested exclusion from the Wayback Machine, or the crawler is blocked by use of a technical flag installed by the site owner called robots.txt, or the site is otherwise inaccessible. When the site is blocked by request or use of a robots.txt flag, the Wayback Machine search engine will indicate this with an error message, such as “blocked site error” or “robots.txt query exclusion error.”

At the inception of the Wayback Machine, the Internet Archive contained 100 terabytes of data that was growing at a rate of 10 terabytes per month. By 2005, the amount of data stored in it is more than a petabyte, with a growth rate of 20 terabytes per month, making the Internet Archive the largest data archive in the world. All of this data is stored in huge server farms in the Presidio of San Francisco.

The archived Web sites are stored across multiple servers. A version of a particular Web site that is shown as indexed on the Wayback Machine may not be available at the time when a user wants to access it. A replica of the Internet Archive is stored at the Bibliotheca Alexandrina in Egypt.¹⁴ If a version of a particular Web site cannot be accessed on the Internet Archives’ primary site, the replica site can be checked.

The replica on the Bibliotheca Alexandrina Web site is not updated frequently, however, and it does not contain as much content. Test searches conducted on *archive.org* reveals many Web sites that do not appear on Alexandrina’s Web site. For example, a search for *cnn.com* yields results for pages from July 2000-September 18, 2001, on the Alexandrina’s Web site, while the *archive.org* site has version from November 26, 2004.

To use the Wayback Machine, users simply go to the *archive.org* Web site, and type in the Internet address¹⁵ in the provided search box. Any versions of the Web site corresponding to the Internet address that are archived on the servers of the Internet Archive will pop up in a chronological list. A user can review this list and select the version or date for review by clicking on the selected date. The archived version of the Web site for the date selected will then appear and can be reviewed.

The nature of the legal dispute may require analysis of multiple archived versions of a particular Web site in order to establish whether and how content changed. For example, in a contract dispute, the question of whether a party offered services or items in violation of terms in the license at issue may require documenting changes in a party’s advertised offerings on its Web site during and after expiration of the license term. Critical text may simply be

eyeballed as part of this analysis to document changes over time. In addition, the Wayback Machine notes changes in an archived Web site with an asterisk. This asterisk system alerts only to changes in text or graphics and not to modifications in internal or external links and or in the source code for the Web site. This may become critical if, for example, the archived Web site is cited as evidence that it was used to link to an offending site. The link to the offending site in the archive version may not, in fact, have existed or existed in the same form at the time that version of the Web site was copied for the archive.

The Wayback Machine also offers a free service of comparing any two versions of an archived site using a technology called DocuComp, which is a patented algorithm licensed by Advanced Software for use in the Wayback Machine. The comparison can show how the contents, including text, images, and links, have changed over time and between any two versions being compared.

“MISSING” ARCHIVED WEB SITES

When a search for an archived Web site has negative results, this does not mean that the Web site does not exist, is not archived, or is only of current vintage. The Web site may have been excluded from the archiving process or in fact, the Web site may be archived but review of the archived versions is blocked. The Internet Archive takes steps to avoid archiving web sites for which the owner has indicated a preference to be excluded. A universal technical standard that indicates an exclusion preference is called the standard for robot exclusion (SRE). A file called robots.txt can be added to the header information on a Web site or specific Web page by an owner, and a denial or disallow command within that file can serve as a flag that the owner does not want the entire Web site or particular Web pages copied or scanned by a Web crawler. In other words, the directions in the robots.txt file can be set to allow full or partial copying or copying exclusion. The Alexa crawler respects this preference and will not copy those sites or pages with a robots.txt file embedded.¹⁶ Alexa Internet and Internet Archive take this respectful technology a step further: When robots.txt is added to a Web site, Alexa will exclude the site from being copied by its crawler, and the Internet Archive will go back into archived sites to remove content already captured.¹⁷

In addition, intellectual property owners who believe that infringing activity is occurring on a Web site may contact the Internet Archive and request exclusion of the offending Web site. The Internet Archive provides specific directions to copyright and trademark owners seeking to have third-party Web sites containing infringing works removed from the archive. These owners must specifically

identify the work allegedly being infringed and where it is located within the Internet Archive collections, contact information, and a statement made under penalty of perjury that use of the work is unauthorized by the copyright owner, along with an electronic or physical signature.¹⁸

The Internet Archives' respect for the exclusion preference of Web site owners and compliance with its own stated policy to remove Web sites with robots.txt flags is the subject of a recent suit in the Eastern District of Pennsylvania brought by Healthcare Advocates against the Internet Archive for, *inter alia*, breach of contract and misrepresentation due to a failure to block access to the plaintiff's archived Web sites.¹⁹ The plaintiff operates a Web site that describes the services of the company, including helping the public get reimbursements for health care expenses, reporting on medical research, providing doctor referrals and information on discount prescriptions and healthcare plans. The company claims copyright in all of the Web site content. In mid-2003, the plaintiff installed the denial text string in the robots.txt file on the computer server hosting its Web site with the expectation that the Internet Archive would prevent users of the Wayback Machine from gaining access to the archived versions of its Web site.

Nevertheless, in another case brought by Healthcare Advocates against a competitor for misuse of proprietary and trade secret information, the defendant's counsel was able to access the archived versions of the plaintiff's Web site on the Wayback Machine by successfully circumventing the security offered by the denial text string in the robots.txt file. This circumvention was apparently facilitated by the fact that "the mechanism preventing www.archive.org from searching a particular web site's host computer server for a denial text string in the robots.txt file more than once per day was 'broken.'" In other words, when the Wayback Machine receives a query for an archived version of a Web site, the Web site is pinged for the presence of a robots.txt file denial string. If the string is found, the query is blocked, but apparently persistent queries will overcome the block. The defendant's counsel in the underlying lawsuit conceded that the plaintiff's archived Web sites on the Wayback Machine had been searched and accessed in connection with that underlying case. That counsel is now co-defendants with the Internet Archive in Healthcare Advocates' suit for copyright infringement and computer hacking.

This lawsuit will test the scope and merits not only of the claims at issue but also the indemnification provision of the Internet Archive's terms of use. Specifically, the terms governing the use of the collection of archived Web pages is predicated on the user's agreement "to indemnify and hold harmless the Internet Archive and its

parents, subsidiaries, affiliates, agents, officers, directors, and employees from and against any and all liability, loss, claims, damages, costs, and/or actions (including attorneys' fees) arising from your use of the Archive's services, the site, or the Collections."²⁰

CAPTURING ARCHIVED WEB SITES

Once an archived Web site has been located, the methods of capturing the virtual pages in a concrete form for use in court can vary. One method is to print each page that appears on the computer screen. The person performing or supervising the search and printing can attest to the date, time, and process used to obtain the printout. This method shows static pages of the Web site without any of the links that may remain active, other than any advertisements pushed to the site, even in the archived state. Similarly, screen-shots of each page viewed can be saved electronically for incorporation into expert reports or affidavits.

Importantly, Internet browsers and specialized tools used by computer forensic experts for downloading Web sites with metadata intact can be used to capture not only the graphical display of a Web page but also the underlying html code that is driving the display. Simply using the file save function on a browser can preserve code that may reveal who authored a contentious Web page. Saving underlying code in the same way may reveal a trademarked name written over and over again in white-on-white text, indicating that it was meant to be revealed to crawling search engines but hidden from a consumer's (or competitor's) naked eye. If two or more archived pages are linked to each other, download tools can provide a fuller layout of a Web site with its underlying code. At trial, this fuller layout can be presented to the judge or jury, and links and related pages can be navigated, much as an historic user might have surfed them.

In addition, specialized software tools are available that allow dynamic presentations, including demonstrations of any link that remains active on the Web site. One such software tool, called Camtasia, can be installed on the computer used to access the archived site to record every keystroke and screen shot appearing during review of the cached Web site. The recording of the review session is documented real time in video-like form that may be stored on a CDR or DVD for submission to court. For example, in a business diversion case, a recording of the cached version of the defendant's prior Web site may be able to show links that remain active and purportedly direct potential customers to the plaintiff's products, but the links instead actually channel users to the defendant's sites.

Beware when capturing an archived Web site that

different browsers display Web sites with differing degrees of accuracy and completeness, and this holds true for archived Web sites and Web pages as well. There are a number of different reasons why some Web pages look different depending on which browser is used to view the page, including browser adherence to Web page standards, browser support of different technologies, and Web sites that do not use Web page standard code. The World Wide Web Consortium (WC3) develops the standard elements for Web site programming, which some browsers adhere to and some do not. For example Firefox and Mozilla adhere to the WC3 standards, while Internet Explorer supports additional non-standard Web-programming technologies. The resulting difference in the way that Web pages are displayed may be as minimal as the color of the scrollbar to as inconvenient as the navigation menus not working or the site content not being displayed at all.

A Web site that uses or requires a certain technology to be viewed will not be displayed correctly or completely by a browser that does not support that technology. For example, Firefox does not support ActiveX, which are software components from Microsoft that enable sound, Java applets, and animations to be integrated in a Web page.²¹ For example, using a browser that supports ActiveX is necessary in order to access the Windows Update Web site, which otherwise will simply not be displayed but with an alert to the viewer that content is hidden from view. The fact that content is not being displayed or displayed in a different way from the original site is not always apparent.

The key to capturing an archived Web site as accurately and completely as possible is to examine the underlying code used to create and support the Web site to determine whether a browser is incompatible. This can be done by an examination of the source code for the initial page of the Web site. The entry point for the Web site usually includes language that will query and collect information from the browser and its computer system settings to determine the best method of providing the information from the site. For Web sites that use only standard html coding, the content and features of the site usually have the least variance across browsers. Where non-standard html coding is revealed, forensic experts capturing Web sites for litigation purposes may display the Web site with multiple browsers as a test to ensure that the display does not vary by browser and if variances are noted, capture the Web site with the browser that displays the most content.

ADMITTING INTERNET ARCHIVE DATA

Information obtained from reputable or government-sponsored online sources has generally been held admis-

sible. For example, in *U.S. Equal Employment Opportunity Commission v. E.I. DuPont De Nemours & Co.*,²² the defendant moved to exclude as an exhibit the printout of a table from the Web site of the US Census Bureau as inadmissible hearsay and lack of trustworthiness. The court denied the motion, stating that the hearsay exception for a public record applied. In addition, the court concluded that the printout was sufficiently authenticated under Federal Rules of Evidence 901(a) since it contained the "internet domain address from which the table was printed, and the date on which it was printed."²³ The court performed its own verification as well, noting that "[t]he Court has accessed the website using the domain address and has verified that the webpage printed exists at that location."²⁴ Similarly, printouts of data from other government-sponsored Web sites have been admitted over objection to the reliability of the information.²⁵

Reported cases involving Web site captures from the Internet Archive are rare, even though *archive.org* is an important resource for litigators trying to establish prior representations or actions on Web sites. Significantly, in the few cases where challenges have been interposed to Internet Archive versions of Web pages, the evidence has been admitted over hearsay and authentication challenges.

The leading case for admission of archived Web sites from the Internet Archive is *Telewizja Polska USA, Inc. v. Echostar Satellite Corporation*.²⁶ The plaintiff in this case claimed that Echostar improperly had used the plaintiff's trademarks in "TV Polonia," a Polish-language television station, to sell subscriptions to the Dish Network satellite TV service after the contract allowing such marketing rights had expired in early 2001. Echostar argued that plaintiff had itself advertised that the Dish Network carried TV Polonia on its Web site after the marketing rights had expired and offered an exhibit of the plaintiff's Web site at various times in 2001 confirming this past Web site content. The plaintiff filed a motion *in limine* to bar Echostar from offering the exhibit on the grounds of double hearsay and lack of authentication. The court rejected these grounds and denied the motion, stating that "the contents of [plaintiff's] website may be considered an admission of a party-opponent and are not barred by the hearsay rule."²⁷ In addition, the court relied on the affidavit of "Ms. Molly Davis, verifying that the Internet Archive Company retrieved copies of the websites as it appeared on the dates in question from its electronic archives."²⁸ The plaintiff "presented no evidence that the Internet Archive is unreliable or biased" or "denied that the exhibit represents the contents of its website on the dates in question" or otherwise "challenged the veracity of the exhibit."

AUTHENTICATION CONSIDERATIONS FOR ARCHIVED WEB SITES

The versions of Web sites and pages archived on Internet Archive can provide valuable and significant probative evidence in a variety of cases. To authenticate copies of prior versions of Web sites obtained from the Wayback Machine, a party proffering the evidence must show, under Federal Rules of Evidence 901(a) that the “matter in question is what its proponent claims.” This can be done by producing the testimony, either orally or in written form, of the person who copied or supervised the copying of the archived Web site and the process followed to accomplish this task. In addition, the proponent must establish the general reliability of the copy.

The capture and use as evidence of archived Web site material must be approached with a full appreciation of three primary technical features and limitations that may affect the archived copy in order to respond to any challenges that may be raised to the completeness, reliability, and authenticity of the copy. For this reason, expertise in digital forensics, including the methods of forensic capture and documentation of the archived Web site proffered, may be recommended depending on the issue for which the archived Web site is being offered.

First, archived Web sites on the Internet Archive are compilations made over time. While the archived versions of Web sites are date- and time-stamped, the pages for each version of the Web site may not have been copied simultaneously. The Alexa crawler may take multiple passes at a Web site over the course of up to two days to try to capture the entire Web site. In short, due to bandwidth and storage constraints, all of the data on a Web site may not be captured at the same time. The Internet Archive explains that “Sites are usually crawled within 24 hours and no more than 48.”²⁹

Second, the archived versions of Web pages available through the Wayback Machine may not contain all of the content on each Web page that is captured. What you see is not always the complete story.

For example, when a Web site contains elements that require interaction with the originating host, copying that page for archiving breaks the necessary link with the original site, thereby reducing the functionality or eliminating entirely that particular element. The result is that the archived Web page or site has missing material, which may not be apparent or flagged for the viewer. Similarly, links originally enabled with a java script, which the Alexa crawl technology disables during the capture of the Web site or Web page, would no longer work.³⁰ The Internet Archive acknowledges: “Not all images are archived nor

are retrievable from the original site. If they no longer exist on the original site then the images will not be available and not displayed within the archived pages.”³¹ Other types of coded content that the crawler technology does not capture include Flash enabled content, some photographic images, and some html coded content.

Moreover, content may not ever be captured if problem technology, such as a password protected pages, or respectful technology, such as a robots.txt flag, is encountered. Additionally, even after the capture is completed, archived copies of Web sites may have content deleted if a robots.txt flag is added to the site or if a request for deletion is sent to the Internet Archive. Thus, the archived copy may show what was captured but not what was skipped or subsequently omitted.

Finally, depending on the technical sophistication of the Web site and its use of internal and outside linked material, the copy of the archived version of the Web site may not show links that existed on the Web site at the time of the original capture. Links that may have worked at the date of capture may be inactive because they simply no longer exist or are not in the archive library.

The links on archived Web sites may remain active but link to different material from that associated with the Web page at the time that it was archived. The linked material may be to current sites or to other stored link sites from a different time. Indeed, links may connect to current active sites and show *current* banner advertisements available at the site, rather than linking to sites as they existed at the date of capture. When the active links on archived Web sites pull information from the current site, the owner of the current Web site can track how many times the Wayback Machine is being queried for archived versions of the Web site. Logs of incoming IP addresses maintained by the server hosting the current Web site can reveal whether the incoming IP address originated with the Internet Archive.³²

Alternatively, the working link may connect to sites or pages archived on the Wayback Machine around the time of the original Web site to which the link connected. The Internet Archive explains: “When you are surfing an incomplete archived site the Wayback Machine will grab the closest available date to the one you are in for the links that are missing. In the event that we do not have the link archived at all, the Wayback Machine will look for the link on the live web and grab it if available.”³³ In short, the process of copying a Web site for archiving may result in changes to the extent that the archived Web site may not show accurately the links that existed at the time shown for the Web site storage date. The Alexa Internet crawler technology rewrites the original link code in html to re-direct links to current or stored links.

Determining whether the content on a linked site is contemporaneous with the archived version of the site or dates from another time may be critical. For example, establishing that a linked promotion to a site containing infringing material persisted after notification from the copyright owner may be important to establish knowledge and intent in a copyright infringement suit. Each link must be checked for the date code embedded in the archived URL, or location within the Wayback Machine database, to verify whether the linked content is contemporaneous, current, earlier or later than the version of the archived Web site or page. The Internet Archive provides the following example: "in this url <http://web.archive.org/web/20000229123340/http://www.yahoo.com/> the date the site was crawled was Feb 29, 2000, at 12:33 and 40 seconds."³⁴

Increasingly, documentation of offending activity that occurred on Web sites of opposing parties is relevant and, in some cases, dispositive of certain types of claims. Searching for, reviewing, and capturing archived copies of Web sites can be easily accomplished from the Internet Archive, but litigators should consider carefully the methods of capture and the issues surrounding the completeness, reliability, and authenticity of the Web site copies.

NOTES

1. See, e.g., *Van Wetrien v. Americontinental Collection Corp.*, 94 F. Supp. 2d 1087, 1109 (D. Or. 2000) (contents of defendant's Web site relevant to determination of whether defendant's conduct was so egregious as to merit an award of punitive damages).
2. The CyberCemetery is located at <http://govinfo.library.unt.edu>.
3. ERC is located at <http://dosfan.lib.uic.edu/ERC/>.
4. The 2004 Presidential Term Web Harvest is located at <http://www.webharvest.gov/collections/peth04/>.
5. <http://www.loc.gov/minerva/collect/elec2002/index.html>.
6. <http://www.loc.gov/minerva/collect/sept11/index.html>.
7. <http://www.loc.gov/minerva/collect/elec2000/index.html>.
8. See <http://searchengineshowdown.com/stats/freshness.shtml>.
9. The Internet Archive is located at www.archive.org.
10. *Kahle v. Ashcroft*, 2004 U.S. Dist. LEXIS 24090, *5 (N.D. Cal. Nov. 19, 2004).
11. <http://pages.alexa.com/company/technology.html>.
12. http://www.archive.org/about/faqs.php#The_Wayback_Machine.
13. <http://websearch.alexa.com/welcome.html>.
14. <http://www.bibalex.org/english/initiatives/internetarchive/web.htm>; see also http://en.wikipedia.org/wiki/Bibliotheca_Alexandrina.
15. The technical term for an Internet address is Universal Resource Locator or URL.
16. Directions for removal of a Web site from the archive are found at <http://www.archive.org/about/exclude.php>.
17. <http://www.archive.org/about/faqs.php#2> ("By placing a simple robots.txt file on your Web server, you can exclude your site from being crawled as well as exclude any historical pages from the Wayback Machine.").
18. *Id.*
19. *Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey*, Civil Action (E.D. Pa., filed July 8, 2005), copy at http://www.geocities.com/ble-drydenet/Healthcare_Advocates_v._Harding_Complaint_FINAL.pdf. Healthcare Advocates, Inc. unsuccessfully moved to have the counts against the law firm for, *inter alia*, violations of the DMCA and the Computer Fraud and Abuse statute added to the underlying complaint, but that motion was denied. *Flynn v. Health Advocate, Inc.*, 2004 U.S. Dist. LEXIS 12536, *12 (E.D. Pa. July 8, 2004).
20. <http://www.archive.org/about/terms.php/>.
21. See <http://webmaster.lycos.co.uk/glossary>.
22. *U.S. Equal Employment Opportunity Commission v. E.I. DuPont de Nemours & Co.*, 2004 U.S. Dist. LEXIS 20753 (E.D. La. Oct. 18, 2004).
23. *Id.* at *5.
24. *Id.*
25. See *Chapman v. San Francisco Newspaper Agency*, 2002 U.S. Dist. LEXIS 18012 at *2 (N.D. Cal. Sept. 20, 2002) (computer printout of page from US Postal Service Web site was sufficiently reliable to be admissible public record). *But see St. Clair v. Johnny's Oyster & Shrimp, Inc.*, 76 F. Supp. 2d 773, 774 (S.D. Tex. 1999) (court deemed plaintiff's proffered data from the US Coast Guard's online vessel database insufficient since "any evidence procured off the Internet is adequate for almost nothing").
26. *Telewizja Polska USA, Inc. v. Echostar Satellite Corp.*, 2004 U.S. Dist. LEXIS 20845 (N.D. Ill.). See also *Attig v. DRG, Inc.*, 2005 U.S. Dist. LEXIS 5183, at *5, n.1 (E.D. Pa. Mar. 30, 2005) (in copyright infringement suit, parties agreed that copies of websites at issue obtained from *archive.org* are admissible evidence); *Louis Vuitton Malletier v. Burlington Coat Factory Warehouse Corp.*, 42 F.3d 532, 535 (2d Cir. 2005) (in trademark infringement suit, evidence of defendant's Web site advertisements presented through *archive.org* capture of the site content at particular time).
27. *Id.* at *16-17.
28. *Id.*
29. http://www.archive.org/about/faqs.php#The_Wayback_Machine.
30. *Id.* ("javascript enabled links and actions are disabled in the comparison results to prevent errant scripts from being run").
31. *Id.*
32. This feature of the Wayback Machine is what alerted Healthcare Advocates in the pending lawsuit discussed *supra*, at n.20 that prior versions of its Web site had not been blocked as requested but instead were being accessed by the defendants.
33. http://www.archive.org/about/faqs.php#The_Wayback_Machine.
34. *Id.*

J. Murphy, N.H. Hashim & P.
O'Connor (2007). "Take me back:
Validating the Wayback Machine."
*Journal of Computer-Mediated
Communication*. 13(1).

Take Me Back: Validating the Wayback Machine

Jamie Murphy

Noor Hazarina Hashim

The University of Western Australia Business School

Peter O'Connor

IMHI, Essec Business School, France

Go to a section in the article:

Abstract

Although fields such as e-commerce, information systems, and computer-mediated communication (CMC) acknowledge the importance of validity, validating research tools or measures in these domains seems the exception rather than the rule. This article extends the concept of validation to one of an emerging genre of web-based tools that provide new measures, the Wayback Machine (WM). Drawing in part on social science tests of validity, the study progresses from testing for and demonstrating the weakest form of validity, face validity, to the more demanding tests for content, predictive, and convergent validity. Finally, the study tests and shows nomological validity, using the diffusion of innovations theory. In line with prior diffusion research, the results of tests for predictive and nomological validity showed significant relationships with organizational characteristics and two WM measures: website age and number of updates. The results help validate these measures and demonstrate the utility of the WM for studying evolving website use.

Introduction

A growing trend is researchers drawing on output from online archival databases such as Google PageRank (Garofalakis, Kappos, & Makris, 2002; Murphy & Scharl, 2007), Google Scholar (Bakkalbasi, Bauer, Glover, & Wang, 2006; Hall, 2006; Kousha & Thelwall, 2007; Pauly & Stergiou, 2005), and several products from Alexa (Palmer, 2002; Ryan, Field, & Olfman, 2003; Thelwall & Wilkinson, 2003; Vaughan & Thelwall, 2003; Veronin, 2002). Most of these studies imply validity, ignore the subject, or note it as a limitation. Yet failure to validate raises issues of trust in research findings (Straub, 1989; Straub, Boudreau, & Gefen, 2004). Computer-mediated communication (CMC) studies validate research instruments such as questionnaires (Koh & Kim, 2003-4; Wade & Nevo, 2005-6), but few address the validity of measures obtained from online tools.

Despite a 1989 call for rigorous instrument validation in management information system research (Straub, 1989), the field has yet "to reach the point where validation is the rule rather than the exception" (Boudreau, Gefen, & Straub, 2001, p. 11). Validation is inadequate, in part due to the difficulty in tracking rapid technological changes (Straub, 1989), yet establishing validity is particularly important for new instruments (Bagozzi, 1981; Hinkin, 1995). In addition to validating research instruments such as survey questions, the computer science field acknowledges validating software or expert systems as an important

step in the development of new tools (Kitchenham, Pfleeger, & Fenton, 1995; O'Leary, Goul, Moffitt, & Radwan, 1990). Similarly, social science often validates research instruments, such as the psychometric properties of questionnaire items (Babbie, 1997; Straub et al., 2004), rather than output from online tools. Validating the output from archival databases is an important new challenge.

In the expert systems domain, a review of validation literature found no standard definition of validity and different terms used interchangeably to describe validity (O'Leary et al., 1990). A business research methods text defines validity as the degree to which a research instrument provides adequate coverage of the topic under study (Sekaran, 2003). In computer science and expert systems, validation is the ability of software or a system to comply with defined standards or adequately represent an expert's knowledge (Kitchenham et al., 1995; Mosqueira-Ray & Moret-Bonillo, 2000; O'Leary et al., 1990).

Common to most definitions across disciplines is determining suitability and accuracy. With regard to types of validity, Straub et al. (2004) argued that *predictive* validity was optional, highly recommended *content* and *nomological* validity, and mandated *convergent* validity. New measures, however, require substantiation of predictive, content, and nomological validity (Bagozzi, 1981; Straub et al., 2004).

Apart from a single study that included convergent and nomological validity for three website measures from Alexa—content, download time, and navigation (Palmer, 2002)—to the authors' knowledge, no CMC studies have validated the output from third-party online tools. Comparing Alexa results with jury ratings and a web-based agent, this sole study found significant correlations and suggested that the three measures had convergent validity. The study further suggested nomological validity for website content and website navigation (Palmer, 2002).

The Wayback Machine (WM) from Alexa provides two complementary measures of website evolution: website age and website updates. Scholars have used the WM to investigate archived website content (Brock, 2005; Hackett & Parmanto, 2005; Ryan et al., 2003; Thelwall & Wilkinson, 2003; Veronin, 2002), infer website age (Vaughan & Thelwall, 2003), and study website evolution (Chu, Leung, Van Hui, & Cheung, 2007). While these studies have drawn upon WM measures, to the authors' knowledge only one published article attempts to validate website content, and no studies have attempted to validate website age. Moreover, no known studies have used a third measure provided by the WM: the number of website updates. Thus, the present study:

1. tests the content validity of three measures provided by the Wayback Machine: archived web pages, website age, and website updates;
2. tests the predictive, nomological, and convergent validity of two measures provided by the Wayback Machine: website age and website updates; and
3. adds to the small number of studies validating measures from third party online tools.

The following sections introduce tests of validity, followed by discussion of the Wayback Machine and diffusion of innovations theory. The article then describes the study population. After testing for face and content validity of three WM measures—website content, website age, and website updates—the article uses the study population to test for predictive, nomological, and convergent validity of the latter two measures. The article closes with suggestions for future research directions for academics studying website evolution or using third-party online tools for research.

Literature Review

Validity Tests

As its name implies, *face validity* relates to face value and relies upon experts' personal opinions and judgment. Because of the vagueness and subjectivity that can result, face validity is a weak test of validity, and some researchers question its use (Sekaran, 2003). Given the lack of validation of third-party online tools, checking face validity seems a reasonable first step prior to moving on to more demanding tests. Closely related to face validity is *content validity*, which ensures that a measure includes an adequate and representative set of items to cover a concept. Content validity also relates to sample-population representativeness, for example, the ability of a questionnaire to represent the larger population. When experts agree that a measure provides adequate coverage of a concept, the measure has content validity (Sekaran, 2003).

Predictive validity, also known as practical or concurrent validity, measures how well an independent variable or set of independent variables relates to the characteristics of research interest (Sekaran, 2003). Scholars debate whether predictive validity falls in the general category of *construct validity* (see below) or the extent that the operationalization of a concept actually measures that concept (Straub, 1989). Predictive validity can also show the applied value of research (Straub et al., 2004). For example, a business could predict its online sales based on the number of website visits and email enquiries. To demonstrate validity, the firm could periodically correlate website visits in a particular month with sales in that or subsequent months. Repeatedly high correlations would suggest predictive validity, thus allowing the firm to use website visits to forecast future sales. Depending on the objective, researchers typically use correlation or regression analyses to test such hypothesized relationships (Hinkin, 1995).

Combined with predictive validity, *nomological* and *convergent* validity help achieve *construct* validity—the empirical and theoretical support for a particular interpretation (Straub, 1989). Nomological, or lawful, validity links a theoretical concept with observable results (Cronbach & Meers, 1955). "If theoretically-derived constructs have been measured with validated instruments and tested against a variety of persons, settings, times, and, in the case of IS research, technologies, then the argument that the constructs themselves are valid becomes more compelling" (Straub et al., 2004, p. 395). *Convergent validity* results when two variables measuring the same construct correlate highly (Straub et al., 2004). Triangulation of multiple research results, rather than relying on a single line of evidence, helps achieve convergent validity.

The Wayback Machine

The Wayback Machine is part of the Internet Archive (www.archive.org), which amasses websites, moving images, texts, audio, and recently, educational resources (FAQs, 2007). Drawing upon results from the Alexa webcrawler, this U.S.-based non-profit organization permanently stores publicly accessible websites in an enormous digital archive. By preserving human knowledge and artifacts and making its collection available to all, the Internet Archive envisions resembling ancient Egypt's legendary Library of Alexandria (FAQs, 2007). The archive contains snapshots of over 55 billion web pages—more information than in any library including the U.S. Library of Congress—even though archiving began only in 1996. The archive adds about 20 terabytes (10^{12} bytes) of digital content monthly (FAQs, 2007), with each sweep of the estimated 16 million archived websites taking over two months (Howell, 2006).

Via the WM, users can view the original version of each site, as well as the dates and content of subsequent updates. To call up archived websites, users type the URL of the desired site into the address box on the WM homepage. The WM then returns the date of original site creation, number and date of site updates, and links to archived sites. Figure 1 shows the WM homepage, and Figure 2 shows the results for a Malaysian hotel, the Timotel in Mersing. The WM also provides information on site updates. An asterisk beside the dates in Figure 2 indicates more than 50% changes to the website since the last visit.

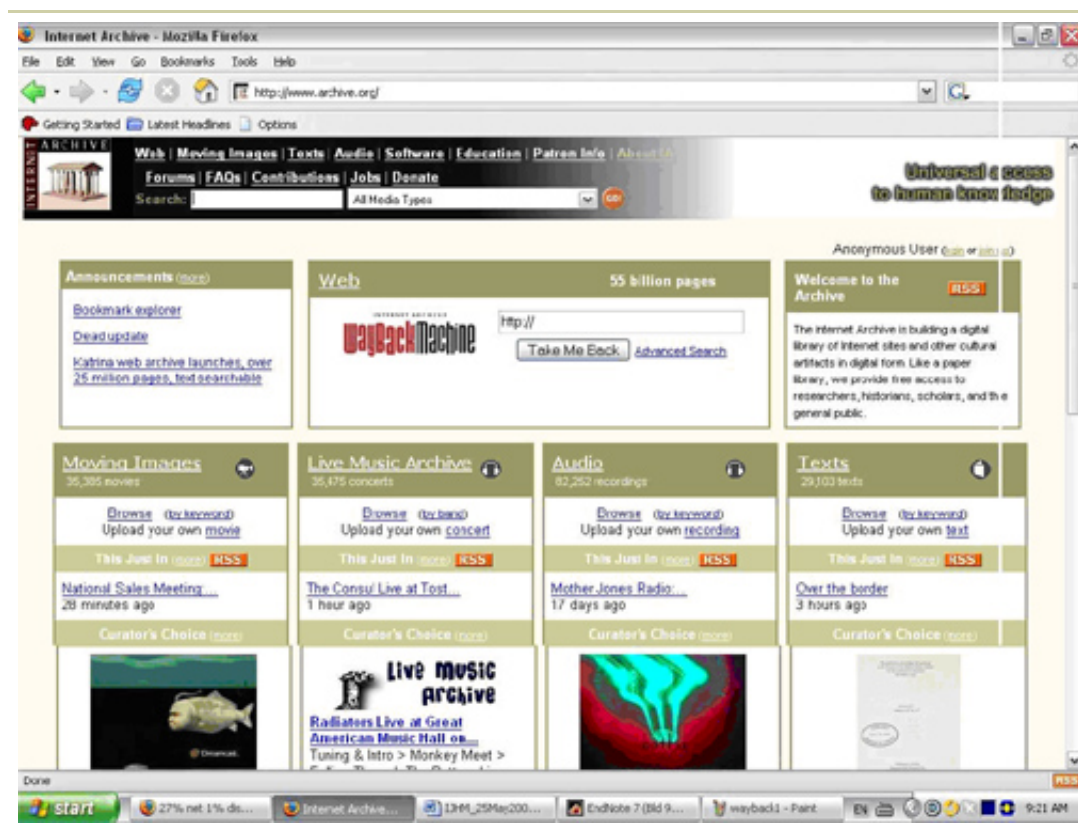


Figure 1. Homepage of the Internet Archive

Search Results for Jan 01, 1996 - May 25, 2006									
1995	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	0 pages	0 pages	1 pages	1 pages	6 pages	11 pages	3 pages	2 pages
				Aug 15, 2000 *	Nov 28, 2001 *	May 24, 2002 *	Feb 03, 2003	Mar 28, 2004	Feb 04, 2005 *
						May 29, 2002	Feb 20, 2003	Jun 02, 2004 *	Mar 05, 2005
						Sep 28, 2002	Apr 21, 2003	Sep 23, 2004 *	
						Sep 30, 2002	Jun 01, 2003		
						Nov 20, 2002	Jun 19, 2003		
						Nov 28, 2002	Jul 23, 2003		
							Aug 08, 2003		
							Oct 10, 2003		
							Oct 25, 2003		
							Nov 18, 2003		
							Dec 18, 2003		

Figure 2. Wayback Machine results for www.timotel.com.my

Tracking the evolution of a site can be useful. For example, a researcher could investigate the evolution of *Hyatt.com*'s online customer relationship programs by analyzing consecutive archived versions of the company's site. As noted earlier, researchers have used the WM to track and measure web content development (Chu et al., 2007; Hackett & Parmanto, 2005). The WM is also gaining legal acceptance with

trademark and intellectual property issues (Howell, 2006). In a landmark 2004 U.S. case, the court ruled that pages culled from the WM were admissible as evidence (Gelman, 2004).

Although massive, the WM has limitations. It archives publicly accessible sites written in simple HTML, but has problems archiving password-protected or dynamic sites (Veronin, 2002). Furthermore, sites can decline inclusion by emailing the Internet Archive or using the Standard for Robot Exclusion (see www.robotstxt.org) to specify files or directories not to crawl (FAQs, 2007). Intellectual property owners concerned about infringements on third party sites can also request removal of such content (FAQs, 2007). Any of these actions stops future indexing, removes site content from the archive, and limits the archives' comprehensiveness. Finally, a condition of use of the Alexa webcrawler is that the Internet Archive must wait at least six months after surveying before including site updates in the archive. Coupled with the requisite time to survey the 55 billion archived pages, this results in a time lag of six to 12 months for an archived snapshot to appear (FAQs, 2007; Howell, 2006).

Diffusion of Innovations

As noted above, testing for nomological validity relies upon an established research stream such as diffusion of innovations (Rogers, 2003). Diffusion studies show that technology evolves from simple to complex use at both the individual (Karahanna, Straub, & Chervany, 1999; Pavlou & Fygenson, 2006) and organizational levels (Cooper & Zmud, 1990; Raho, Belohlav, & Fiedler, 1987; Zmud & Apple, 1992).

Since 1994, academics have investigated evolutionary aspects of business Internet use (Teo & Pian, 2004) and noted the importance of understanding the evolution of websites (Hoffman, Novak, & Chatterjee, 1996; Park & Thelwall, 2003). For example, net-based customer service may evolve over five phases—experimentation, value creation, focus, differentiation, and relationships (Piccoli, Brohman, Watson, & Parasuraman, 2004). Alternatively, a longitudinal study proposed that e-commerce websites evolved over four eras, from the pre-web to the integrative web era (Chu et al., 2007).

Examining websites' evolutionary aspects helps researchers investigate what factors lead to successful website implementation, including which features organizations add, and leave, on their websites. Evolution itself, however, is a research limitation; a single evaluation at a single point in time cannot capture such evolution. While longitudinal studies would let researchers track changing relationships, performing multiple evaluations is difficult and cumbersome (Chatterjee, Grewal, & Sambamurthy, 2002). Furthermore, some websites may no longer exist and some changes are ephemeral. For instance, a study of over 1,000 websites across six categories found only two-thirds of the sites still functioning at the same URL five years later (McMillan, 2002).

Another research limitation of diffusion studies is relying upon stated behavior rather than measuring actual behavior (Damanpour, 1991; Rogers, 2003). For example, to measure website age, researchers could email webmasters to ask when their websites first went online. However, a webmaster might not reply, might not know, or might give incorrect information. Domain name age, based on when an organization originally registered its domain name—such as Hyatt hotels registering *Hyatt.com*—provides an actual measure of Internet adoption (Adamic & Huberman, 2000; Murphy, Olaru, & Schegg, 2006). Yet domain name age as a measure of website evolution has limitations. With names registered in the most common domain, *.com*, changes in domain name registrars render the recorded age invalid (Murphy et al., 2006). Similarly, organizations may buy a domain name but wait months or years before hosting a website at that name, thus making the registration date an unreliable measure of when a website went live. Using data from the WM, which archives actual website pages, helps overcome such limitations and establish the real date of site creation.

The study context is Malaysian hotels, for three reasons. Of those industries going online, travel leads other service industries in its share of e-commerce (Dinlersoz & Hernández-Murillo, 2005). Second, hospitality e-commerce studies often draw upon the diffusion of innovations (Matzler, Pechlaner, Abfalter, & Wolf, 2005; Murphy et al., 2006; Murphy, Olaru, Schegg, & Frey, 2003; Wang & Fesenmaier, 2005), and using this research stream facilitates testing for nomological validity. Finally, most tourism Internet research focuses on developed countries; there is a lack of research in developing countries (Frew, 2000; Hashim, Murphy, & Hashim, 2007). This may be partly due to Internet use being at an early and formative phase in developing countries such as Malaysia (Le & Koh, 2002). Studies of Malaysia's hospitality industry are limited, particularly regarding Internet use (Hashim & Murphy, 2007). Using Malaysian hotels contributes to the body of knowledge in these domains, while at the same time achieving the study objectives.

Data Preparation and Preliminary Nomological Results

Testing the content, predictive, convergent, and nomological validity of the WM measures necessitated a database. With no comprehensive list of Malaysian hotel websites available, the study began with 540 hotels registered with Malaysia's Ministry of Tourism, and the Malaysian Accommodation Directory's (MAD) 2003/2004 list of hotel website addresses. In May 2006, keying the 540 hotels' names into Google and Yahoo! helped find more hotel websites and verify the MAD website addresses, yielding 310 websites. The WM failed to give results for 19 sites (about 6%), due to trouble locating the site or the site declining indexing by the Internet Archive. Of the remaining 291 websites, some chain hotels shared the same domain name, such as *hyatt.com* and *hilton.com* for all Hyatt and Hilton hotels in Malaysia. To avoid duplication, excluding 116 hotels with the same domain name left 175 websites. Of these 175 hotels, 96 hotels hosted their website in the global *.com* domain, and 79 hosted their website in Malaysia's country domain, *.my*.

Diffusion of Innovations Findings

Table 1 shows the final sample and suggests that in line with diffusion of innovations research, high rated, chain-affiliated, and large hotels tended to lead in website adoption (Murphy et al., 2003; Siguaw, Enz, & Namiasivayam, 2000; Wei, Ruys, van Hoof, & Combrink, 2001). The first five-star hotel went online almost three years earlier than the first one-star hotel, early 1997 versus late 2000. The first chain hotel went online nearly a year earlier than the first non-affiliated hotel, late 1996 versus mid 1997. Finally, the first online hotel with over 300 rooms was about two years older than the first online hotel with under 200 rooms.

	Websites accessible via the WM	Sample without same domain name	Sample with <i>.my</i> domain	First website	Most updates from 1996-2005
Rating					
1-star	4	3	0	10.11.2000	35
2-star	50	31	12	3.11.1999	35
3-star	95	62	28	27.8.1997	63
4-star	71	48	24	22.12.1996	72
5-star	71	30	15	25.1.1997	60
Affiliation					
Chain	205	89	40	22.12.1996	72
Non-chain	86	86	39	27.8.1997	63
No. of					

Rooms					
1-99	70	53	16	1.12.1998	63
100-199	84	50	21	25.1.1997	33
200-299	55	35	22	25.1.1998	56
>299	82	37	20	22.12.1996	72
Total	291	175	79		

Table 1. Sample characteristics

Similarly, high rated, chain-affiliated, and large hotels led in updating their websites. The five-star hotel with the most updates from 1996-2006 changed its site 60 times, compared to 35 times for the leading one-star hotel. Likewise, the leading chain-affiliated large hotel, which was also a large hotel, made 72 updates on its website versus 63 updates for the leading non-affiliated hotel that was also a small hotel. This discussion of website age and number of updates suggests *nomological validity* in line with the diffusion of innovations, but the results are just for one hotel—the leading hotel in each category—and not the entire sample of hotels.

Thus, the next section tests the validity of the Wayback Machine's website age and website updates using the entire sample. Three transformations were necessary prior to testing. A new variable, update frequency, was the website age divided by the total number of website updates. Using this new variable, the most frequently updated website was a three-star independent hotel in Terengganu, which averaged an update every 35 days. At the other extreme, a two-star independent hotel in Melaka updated its website on average once every five years. As update frequency and the number of rooms had an abnormal distribution based on a one-sample Kolmogorov-Smirnoff test, a logarithmic function transformed these two variables into a normal distribution.

Validating the Wayback Machine

The following discussion draws on instrument validation and research of individual measures to validate three measures provided by the WM—website content, website age, and website updates. Starting with the weaker and more subjective tests, this study assessed *face validity* based on published research, feedback from three website managers, and a comparison with Malaysia's domain name database. The courtroom acceptance of the WM (Gelman, 2004; Howell, 2006), mentioned earlier in this article, demonstrates face validity by legal experts. Next, an email invited two Malaysian hoteliers to test their website in the WM. The WM provided archived versions of their sites, and they agreed that the WM provided accurate ages and archived versions. Similarly, an author of this study verified that the WM provided accurate dates and versions of one U.S. and four Australian websites that he managed. The study also examined the face validity of the WM by investigating four hotel homepages shown in a 1996 study (Murphy, Forrest, Wotring, & Brymer). The WM results showed the same homepages as those in the article.

A final test of face validity compared the website age provided by the Wayback Machine with the domain name age provided by Mynic, Malaysia's domain name registrar (whois.mynic.net.my). In principle, a hotel would register a domain name to house the website prior to launching the website. Comparing the WM website age with the domain name age for the 79 hotels using a .my domain name showed that 68 hotels had a domain name age older than the WM website age. Three hotels changed domain names, evidenced by the links and content on archived web pages. For example, the Hotel Flamingo began at www.twosteps.com/flamingo on August 23, 2000 and then changed to www.flamingo.com.my on June 3, 2002. The other eight hotels changed their Mynic information, resetting the registration date on file with

Mynic. These two issues highlight shortcomings of using domain name age as a measure of Internet adoption and provide face validity for the Malaysian hotels' website age.

Content validity was assessed based on the representativeness of websites and adequacy of the website age information provided by WM. As noted above, the WM provided universal coverage for the four sites in the published study and the seven sites managed by three webmasters. Furthermore, as noted in the data preparation section, the WM returned archived versions for 291 of 310 hotel websites, which suggests representativeness. In summary, confirmation by website managers, comparison with a published study, and representation of 291 Malaysian hotels in the WM suggest face and content validity of the WM's website age, website updates, and archived web pages.

Predictive validity stemmed from the number of website updates recorded. Literature on the evolutionary nature of websites (Chatterjee et al., 2002; Chu et al., 2007; Piccoli et al., 2004; Teo & Pian, 2004) led to the prediction that older websites would have a higher average frequency of updates. The result of a one-tailed Pearson correlation test—a significant positive relationship between website age and the logarithmic value of update frequency ($r=.274$, $n=175$, $p<.001$)—shows older websites were updated more frequently and suggests predictive validity.

The diffusion of innovations served as the theoretical base for testing *nomological validity*. This theory argues that certain organizational characteristics relate positively to organizational technology use (Matzler et al., 2005; Wang & Fesenmaier, 2005). U.S. and Swiss studies showed that high rated, large, and affiliated hotels led in technology adoption (Murphy et al., 2003; Siguaw et al., 2000). Compared to lower rated, smaller, or non-affiliated hotels, such hotels had more resources and expertise to facilitate IT implementation. Similarly, emerging Malaysian research and early global research found that large, high rated, and affiliated hotels led in the use of advanced website features (Hashim & Murphy, 2007; Wei et al., 2001). Based on the similarity in these studies, star rating, hotel size, and brand affiliation were the independent variables for testing nomological validity.

Table 2 shows the results of one-way Pearson correlation tests for the logarithmic number of rooms, Spearman correlation tests for star rating, and independent t-tests for chain-affiliation against the dependent variables of website age and number of updates. As mentioned earlier, the analysis used logarithmic values for the update frequency and number of rooms. Given the possible correlation among the three independent variables—size, number of stars, and affiliation—two multiple regression tests examined the predictive importance of the independent variables on website age and number of updates. No independent variables were significant predictors for number of updates, and star rating was a significant predictor of website age ($\beta=.203$, $p=.031$).

Correlation coefficient/ t-value, significance level	Size	Rating	Affiliation
Website age in days	0.161, $p=0.017$	0.239, $p=0.001$	2.737, $p=0.004$
Average update frequency	0.112, $p=0.070$	0.193, $p=0.005$	1.775, $p=0.039$

Table 2. Correlation and T-test results for website age and number of updates (N=175)

Although the low correlation coefficients in Table 2 indicated significant relationships, and the multiple regressions showed low predictive importance, the results were in line with diffusion of innovations research. Larger, higher-rated, and affiliated hotels launched their websites earlier and updated their websites more often than smaller, lower-rated, and non-affiliated hotels did, helping support nomological

validity.

Convergent validity was evaluated by measuring the relationship between domain name age and the creation date of a website at that address. Despite the limitation of a temporal gap between owning a domain name and having a live website, studies use an organization's domain name age as a proxy for Internet adoption (Adamic & Huberman, 2000; Murphy et al., 2006). Although as explained earlier, a domain name age is an imperfect proxy, a high positive correlation between a website's domain name age and that same website's age as provided by the WM would suggest convergent validity.

Establishing the age of names in global domains such as *.com* or *.org*, however, is problematic. On November 30, 1999, the Internet Corporation for Assigned Names and Numbers shifted from a sole domain registrar to a Shared Registration System (SRS) with multiple registrars in the *.com*, *.net*, and *.org* domains (see www.icann.org for a history of domain names). The SRS makes gathering valid global domain name ages unreliable, as companies may change domain registrars, resetting their domain name's creation date and rendering the data invalid (Murphy et al., 2006).

At the country level, however, such as *.at* and *.my* for Austria and Malaysia respectively, gathering the domain name age is less problematic. There is usually just one domain name database for each country, such as in Malaysia. Due to the difficulty validating ages in the *.com* domain, the study used the 79 websites with a *.my* domain to test convergent validity. Eliminating the 11 hotels that changed domains or Mynic information, the result of one-way Pearson correlation for the 68 hotels hosted in *.my* showed a significant positive correlation between website age and domain name age ($r=.933$, $p<.001$). This strong correlation supports convergent validity for the website age provided by the Wayback Machine.

Conclusions and Future Research

Researchers frequently adopt instruments from other studies, which can contribute to flawed measures for at least two reasons. Researchers fail to validate the adopted instrument or make major alterations to a validated instrument without re-testing it (Straub, 1989). This study reinforces the importance of the first reason, failure to validate, for metrics from the growing field of third-party tools such as those provided by Google and Alexa. As researchers continue to use these tools, it is important to address the validity of both the tools and their measures.

This article augments research on the evolutionary and dynamic nature of CMC (Chatterjee et al., 2002; Chu et al., 2007; Hoffman et al., 1996; Murphy et al., 2006; Park & Thelwall, 2003; Piccoli et al., 2004; Teo & Pian, 2004) by suggesting and validating two hitherto underutilized website adoption measures, website age and number of website updates. Website age helps overcome limitations associated with domain name age and gives researchers a valid temporal measure of website adoption. Furthermore, the archived websites and number of website updates allow researchers and practitioners to study websites over time.

Although the Wayback Machine has limitations such as not indexing some websites, the results of this study showed content validity for three WM measures—website content, website age, and number of updates—as well as predictive, nomological, and convergent validity for website age and number of website updates. This article thus adds to the minimal research on validating online third party tools. Validation studies often deal with a survey instrument or a software process, but results from third party tools such as Alexa seem a new and fruitful area for validation studies.

Future Research

As this study investigated just four types of validation, future research should address other validation tests, as well as the reliability of third party tools (Boudreau et al., 2001; Straub, 1989; Straub et al., 2004). While this article suggests that the WM provides valid website ages and website updates, future research should revisit these two WM metrics in other industries and extend the concept of validation to measures from other web-based third party tools. For example, Alexa provides measures of website popularity and incoming links to a website. Google provides a toolbar that ranks websites on Google's proprietary PageRank, and a beta tool, Google Scholar (scholar.google.com), provides popularity measures for scholarly articles (Bakkalbasi et al., 2006; Hall, 2006; Jacsó, 2005, 2006; Kousha & Thelwall, 2007; Pauly & Stergiou, 2005). While widely used, to the authors' knowledge these tools remain unvalidated.

In addition, this article relied on social science methods to validate measures provided by the WM, rather than validating the WM itself. CMC researchers should draw upon and collaborate with colleagues in computer science and expert systems to apply methods such as evaluation, validation, and verification to the WM (Kitchenham et al., 1995; Mosqueira-Ray & Moret-Bonillo, 2000; O'Leary et al., 1990).

Finally, now that the Wayback Machine seems validated as a viable research tool, an interesting range of research possibilities arise. Researchers can now have greater confidence in the data generated by the tool and can incorporate such data into their research on website development and e-commerce. As suggested elsewhere in this article, the WM facilitates studies of website development over time. Taking a historical perspective and exploiting this opportunity should lead to a better understanding of website evolutions in domains such as e-commerce and Web 2.0.

Acknowledgment

The authors presented an earlier and abridged version of this manuscript at the January 2007 ENTER Conference in Ljubljana, Slovenia.

References

- Adamic, L. A., & Huberman, B. A. (2000). The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce*, 1 (1), 5-12.
- Babbie, E. R. (1997). *The Practice of Social Research* (8th ed.). Belmont, CA: Wadsworth Publishing.
- Bagozzi, R. P. (1981). An examination of the validity of two models of attitude. *Multivariate Behavioral Research*, 16 (3), 323-359.
- Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus, and Web of Science. *Biomedical Digital Libraries*, 3 (7). Retrieved October 20, 2007 from <http://www.bio-diglib.com/content/3/1/7>
- Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly*, 25 (1), 1-16.
- Brock, A. (2005). A belief in humanity is a belief in colored men: Using culture to span the digital divide. *Journal of Computer Mediated Communication*, 11 (1), article 17. Retrieved September 18, 2007 from <http://jcmc.indiana.edu/vol11/issue11/brock.html>
- Chatterjee, D., Grewal, R., & Sambamurthy, V. (2002). Shaping up for e-commerce: Institutional enablers of the organizational assimilation of web technologies. *MIS Quarterly*, 26 (2), 65-89.

- Chu, S.-C., Leung, L. C., Van Hui, Y., & Cheung, W. (2007). Evolution of e-commerce web sites: A conceptual framework and a longitudinal study. *Information and Management*, 44 (2), 154-164.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychology Bulletin*, 52 (4), 281-302.
- Cooper, R. B., & Zmud, R. W. (1990). Information technology implementation research: A technological diffusion approach. *Management Science*, 36 (2), 123-139.
- Damanpour, F. (1991). Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal*, 34 (3), 555-590.
- Dinlersoz, E. M., & Hernández-Murillo, R. (2005). The diffusion of electronic business in the United States. *Federal Reserve Bank of St. Louis Review*, 87 (1), 11-34.
- FAQs. (2007). *The Wayback Machine: Frequently asked questions*. Retrieved April 5, 2007 from <http://www.archive.org/about/faqs.php>
- Frew, A. J. (2000). Information technology and tourism: A research agenda. *Information Technology and Tourism*, 3 (2), 99-110.
- Garofalakis, J. G., Kappos, P., & Makris, C. (2002). Improving the performance of web access by bridging global ranking with local page popularity metrics. *Internet Research: Electronic Networking Applications and Policy*, 12 (1), 43-54.
- Gelman, L. (2004). Internet archive's web page snapshots held admissible as evidence. *Packets*, 2 (3). Retrieved October 20, 2007 from <http://cyberlaw.stanford.edu/packets002728.shtml>
- Hackett, S., & Parmanto, B. (2005). A longitudinal evaluation of accessibility: Higher education web sites. *Internet Research*, 15 (3), 281-294.
- Hall, C. M. (2006). The impact of tourism knowledge: Google Scholar, citations, and the opening up of academic space. *e-Review of Tourism Research*, 4 (5), 119-136.
- Hashim, N. H., & Murphy, J. (2007). Branding on the web: Evolving domain name usage among Malaysian hotels. *Tourism Management*, 28 (2), 621-624.
- Hashim, N. H., Murphy, J., & Hashim, N. M. (2007). Islam and online imagery on Malaysian tourist destination websites. *Journal of Computer Mediated Communication*, 12 (3), article 16. Retrieved September 18, 2007 from <http://jcmc.indiana.edu/vol12/issue13/hashim.html>
- Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21 (5), 967-988.
- Hoffman, D. L., Novak, T. P., & Chatterjee, P. (1996). Commercial scenarios for the web: Opportunities and challenges. *Journal of Computer Mediated Communication*, 1 (3). Retrieved October 20, 2007 from <http://jcmc.indiana.edu/vol1/issue3/hoffman.html>
- Howell, B. A. (2006). Proving web history: How to use the Internet archive. *Journal of Internet Law*, 9 (8), 3-9.
- Jacsó, P. (2005, December). *Comparison and analysis of the citedness scores in Web of Science and Google Scholar*. Paper presented at the Proceedings of Digital Libraries: Implementing Strategies and Sharing Experiences: 8th International Conference on Asian Digital Libraries, ICADL, Bangkok, Thailand.

- Jacsó, P. (2006). Google Scholar: The pros and the cons. *Online Information Review*, 29 (2), 208-214.
- Karahanna, E., Straub, D. W., & Chervany, N. L. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly*, 23 (2), 183-213.
- Kitchenham, B., Pfleeger, S. L., & Fenton, N. (1995). Towards a framework for software measurement validation. *IEEE Transactions on Software Engineering*, 21 (12), 929-944.
- Koh, J., & Kim, Y.-G. (2003-4). Sense of virtual community: A conceptual framework and empirical validation. *International Journal of Electronic Commerce*, 8 (2), 75-93.
- Kousha, K., & Thelwall, M. (2007, in press). Google Scholar citations and Google web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, 58 (7).
- Le, T. T., & Koh, A. C. (2002). A managerial perspective on electronic commerce development in Malaysia. *Electronic Commerce Research*, 2 (1/2), 7-29.
- Matzler, K., Pechlaner, H., Abfalter, D., & Wolf, M. (2005). Determinants of response to customer e-mail enquiries to hotels: Evidence from Austria. *Tourism Management*, 26 (2), 249-259.
- McMillan, S. J. (2002). Longevity of websites and interactive advertising communication. *Journal of Interactive Advertising*, 2 (2). Retrieved October 20, 2007 from <http://www.jiad.org/vol2/no2/mcmillan/>
- Mosqueira-Ray, E., & Moret-Bonillo, V. (2000). Validation of intelligent systems: A critical study and a tool. *Expert Systems with Applications*, 18 (1), 1-16.
- Murphy, J., Forrest, E. J., Wotring, C. E., & Brymer, R. A. (1996). Hotel management and marketing on the Internet. *Cornell Hotel and Restaurant Administration Quarterly*, 37 (3), 70-82.
- Murphy, J., Olaru, D., & Schegg, R. (2006). Investigating the evolution of hotel Internet adoption. *Information Technology and Tourism*, 8 (3/4), 161-178.
- Murphy, J., Olaru, D., Schegg, R., & Frey, S. (2003). The bandwagon effect: Swiss hotels' website and e-mail management. *Cornell Hotel and Restaurant Administration Quarterly*, 44 (1), 71-87.
- Murphy, J., & Scharl, A. (2007). An investigation of global versus local online branding. *International Marketing Review*, 24 (3), 297-312.
- O'Leary, T. J., Goul, M., Moffitt, K. E., & Radwan, A. E. (1990). Validating expert systems. *IEEE Intelligent Systems*, 5 (3), 51-58.
- Palmer, J. W. (2002). Web site usability, design, and performance metrics. *Information Systems Research*, 13 (2), 151-167.
- Park, H. W., & Thelwall, M. (2003). Hyperlink analysis of the World Wide Web: A review. *Journal of Computer-Mediated Communication*, 8 (4). Retrieved October 20, 2007 from <http://jcmc.indiana.edu/vol8/issue4/park.html>
- Pauly, D., & Stergiou, K. I. (2005). Equivalence of results from two citation analyses: Thomson ISI's citation index and Google's scholar service. *Ethics in Science and Environmental Politics*, 2005, 33-35. Retrieved October 20, 2007 from <http://www.int-res.com/articles/esep/2005/E65.pdf>
- Pavlou, P. A., & Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: An

extension of the theory of planned behavior. *MIS Quarterly*, 30 (1), 115-143.

Piccoli, G., Brohman, M. K., Watson, R. T., & Parasuraman, A. (2004). Net-based customer service systems: Evolution and revolution in web site functionalities. *Decision Sciences*, 35 (3), 423-455.

Raho, L. E., Belohlav, J. A., & Fiedler, K. D. (1987). Assimilating new technology into the organization: An assessment of McFarlan and McKenney's model. *MIS Quarterly*, 11 (1), 47-57.

Rogers, E. M. (2003). *Diffusion of Innovations* (5th ed.). New York: Simon & Schuster.

Ryan, T., Field, R. H. G., & Olfman, L. (2003). The evolution of U.S. state government home pages from 1997 to 2002. *International Journal of Human-Computer Studies*, 59 (4), 403-430.

Sekaran, U. (2003). *Research Methods for Business: A Skill Building Approach* (4th ed.). New York: John Wiley & Sons Inc.

Siguaw, J. A.,ENZ, C. A., & Namiasivayam, K. (2000). Adoption of information technology in U.S. hotels: Strategically driven objectives. *Journal of Travel Research*, 39 (November), 192-201.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly*, 13 (2), 147-169.

Straub, D. W., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems*, 13, 380-427.

Teo, T., & Pian, Y. (2004). A model for web adoption. *Information and Management*, 41 (4), 457-468.

Thelwall, M., & Wilkinson, D. (2003). Three target document range metrics for university websites. *Journal of the American Society for Information Science and Technology*, 54 (6), 490-497.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology*, 54 (1), 29-38.

Veronin, M. A. (2002). Where are they now? A case study of health-related web site attrition. *Journal of Medical Internet Research*, 4 (2). Retrieved October 20, 2007 from <http://www.jmir.org/2002/2002/e2010/>

Wade, M. R., & Nevo, S. (2005-6). Development and validation of a perceptual instrument to measure e-commerce performance. *International Journal of Electronic Commerce*, 10 (2), 123-146.

Wang, Y., & Fesenmaier, D. R. (2005). Identifying the success factors of web-based marketing strategy: An investigation of convention and visitors bureaus in the United States. *Journal of Travel Research*, 43 (3), 1-11.

Wei, S., Ruys, H. F., van Hoof, H. B., & Combrink, T. E. (2001). Uses of the Internet in the global hotel industry. *Journal of Business Research*, 54 (3), 235-241.

Zmud, R. W., & Apple, L. E. (1992). Measuring technology incorporation/infusion. *Journal of Product Innovation Management*, 9 (June), 148-155.

About the Authors

Associate Professor [Jamie Murphy's](#) background includes complementary industry and academic experience. In addition to owning/managing hospitality businesses, he served as the European Marketing Manager for U.S. sports companies. Dr. Murphy's research focus is effective use of the Internet for citizens, businesses, and governments.

Address: The University of Western Australia Business School, Crawley, WA 6009, Australia

[Noor Hazarina Hashim](#) is a Ph.D. candidate in Internet Marketing at the School of Economics and Commerce, University of Western Australia. In Malaysia, she lectures in marketing at the University of Technology Malaysia. Her research interests include the evolution of website and email, and effective Internet marketing.

Address: The University of Western Australia Business School, Crawley, WA 6009, Australia

[Peter O'Connor](#) is Professor of Information Systems at Essec Business School France and serves as Academic Director of Institute de Management Hotelier International (IMHI), its specialized MBA program in hospitality management. His primary research, teaching, and consulting interests focus on the use of information technology in the hospitality sector. Previously he held a visiting position at the Cornell School of Hotel Administration and worked in a variety of international positions in hospitality management in sectors ranging from luxury hotels to contract food services.

Address: IMHI, Essec Business School, 95021 Cergy-Pontoise Cedex, France

S. Schneider and K. Foot (2004). "The web as an object of study." *New Media and Society*. 6(1): 114-122.



The web as an object of study

STEVEN M. SCHNEIDER
SUNY Institute of Technology, USA

KIRSTEN A. FOOT
University of Washington, USA

INTRODUCTION

Over the last decade, as email, the world wide web and various digital technologies have emerged, scholars of new media have employed a variety of methodological strategies to explore the social, political and cultural phenomena associated with the growth of these applications. Several recently-published edited volumes highlight the range of methods employed in research regarding social dimensions of internet technologies (Gauntlett, 2000; Howard and Jones, 2003; Jones, 1999; Mann and Stewart, 2000). These collections, along with recent issues of scholarly journals, demonstrate that traditional methods of social research, such as ethnography (e.g. Hakken, 1999; Hine, 2000; Markham, 1998), textual analysis (e.g. Crowston and Williams, 2000; Mitra, 1999; Mitra and Cohen, 1999), focus groups (e.g. Price and Capella, 2001, Stromer-Galley and Foot, 2002), surveys (e.g. Parks and Floyd, 1996; Schmidt, 1997; Smith, 1997; Yun and Trumbo, 2000) and experiments (e.g. Iyengar, 2002) have been adapted for use online in order to investigate both online and offline phenomena. In addition, some scholars have found it useful to employ internet applications as bases for studies of purely offline phenomena (e.g. Witte et al., 2000). However, our focus is on the development of methods for studying the social dimensions of the internet itself, and in particular, the web.

As the web has emerged as a distinct media form in the past 10 years, it has been viewed increasingly as an object of study by social researchers. The ongoing evolution of the web poses challenges for scholars as they seek to develop methodological approaches that permit robust examination of web

phenomena. Some of these challenges stem from the nature of the web, which is a unique mixture of the ephemeral and the permanent. There are two aspects to the ephemerality of web content. First, web content is ephemeral in its transience, as it can be expected to last for only a relatively brief time. From the perspective of the user or visitor (or researcher), specialized tools and techniques are required to ensure that content can be viewed again at a later time. Second, web content is ephemeral in its construction – like television, radio, theater, and other ‘performance media’ (Hecht et al., 1993; Stowkowski, 2002). Web content, once presented, needs to be reconstructed or represented in order for others to experience it. Although webpages are routinely reconstructed by computers without human intervention (when a request is forwarded to a web server), it nevertheless requires some action by the producer (or the producer’s server) in order for the content to be viewed again. In other words, the experience of the web, as well as the bits used to produce the content, must be intentionally preserved in order for it to be reproduced (Arms et al., 2001). Older media – including printed materials, film and sound recordings, for example – can be archived in the form in which they are presented; no additional steps are needed to recreate the experience of the original.

At the same time, the web has a sense of permanence that clearly distinguishes it from performance media. Unlike theater, or live television or radio, web content must exist in a permanent form in order to be transmitted. The web shares this characteristic with other forms of media such as film, print, and sound recordings. However, the permanence of the web is somewhat fleeting. Unlike any other permanent media, a website may destroy its predecessor regularly and procedurally each time it is updated by its producer; that is, absent specific arrangements to the contrary, each previous edition of a website may be erased as a new version is produced. By analogy, it would be as if each day’s newspaper was printed on the same piece of paper, obliterating yesterday’s news in order to produce today’s.

The ephemerality of the web requires that proactive steps be taken in order to allow a recreation of web experience for future analyses. The permanence of the web makes this eminently possible. Although saving websites is not as easy as, for example, saving editions of a magazine, archiving techniques are evolving in such a way as to facilitate scholarly research of websites. As distinct from other ephemeral media, the web can be preserved in nearly the same form as it was originally ‘performed’ (Kahle, 1997; Lyman, 2002; Lyman and Kahle, 1998) and analyzed at a later time. Web archiving enables more rigorous and verifiable research, as well as developmental analyses that are time sensitive (e.g. Foot et al., 2003).

APPROACHES IN WEB STUDIES

Some of the broad questions currently under investigation by web scholars include the following.

- What forms of communicative actions are being inscribed on the web, and how do they change over time?
- How do the actions of web producers enable and/or constrain the potential actions of web users?
- What kinds of user experiences are potentiated on, and between, particular websites?
- How are relations between web producers, as well as between producers and users, enacted and mediated via web texts and links?

These kinds of research questions, along with the increasingly complex web applications that are altering traditional relationships between media form and content, challenge traditional approaches to social research. Web-based media require new methods of analyzing form and content, along with processes and patterns of production, distribution, usage and interpretation.

We identify three sets of approaches that have been employed in web-related research over the last decade. These approaches are not necessarily mutually exclusive, and some studies cited below employed more than one approach. Distinguishing between these approaches helps to establish the trajectory of web studies; highlighting the strengths and weaknesses of each focuses attention on the methodological challenges that are associated with the field of web studies.

The first set of approaches that we identify employ *discursive* or *rhetorical* analyses of websites; it is more concerned with the content of a website than its structuring elements. Studies employing these approaches focus on the texts and images that are contained on webpages, and/or on webpages/websites as texts in a Foucauldian sense (e.g. Baym, 1999; Benoit and Benoit, 2000; Sillaman, 2000; Warnick, 1998). Studies using a discursive/rhetorical approach, especially those that take broad views of what constitutes text, contribute significantly to our understanding of communicative phenomena on the web. However, we contend that the classic arguments regarding the inseparability of form and content in traditional media are especially applicable to the web, and thus studies of web 'content' that overlook the structuring elements of a webpage or site are also limited. Another limitation within this set of approaches is the paucity of analytical tools for making sense of the links among webpages and between websites. Some studies of hypertext intertextuality include analyses of cross-site linking, (e.g. Mitra, 1999; Warnick, 2001), but most content-focused studies of the web tend to reflect and perpetuate what we

believe is an inadequate construction of the web as merely a collection of texts. As Burbules and Callister (2000: 83) observe, 'people usually see points or texts as primary, and the links between them as mere connectives'. We agree with their claim that links are 'associative relations that change, redefine and provide enhanced or restricted access to the information they comprise', and we support the argument offered by Berners-Lee (2000), Mitra (1999), Odlyzko (2001) and others that, on the web, connectivity matters as much as content.

We characterize the second set of approaches as *structural* or *feature* analyses. Studies in this genre tend to use individual websites as their unit of analysis, focusing on the structure of the site, such as the number of pages and their hierarchical ordering, or on the features found on the pages within the site, for example, the presence of a search engine, privacy policy, or multiple navigation options (Benoit and Benoit, 2000; D'Alessio, 1997, 2000; Hansen, 2000; McMillan, 1999). Although understanding the structural and feature aspects of a particular site is important, our primary concern with these approaches is that they do not afford systematic analysis of an individual site's situatedness in the larger web, that is, the external pages to which it links and are linked to it. Another type of structural analysis employs computer-assisted, macro-level network analysis methods for mapping linking patterns (e.g. Jackson, 1997; Park, 2003; Park and Thelwall, 2003; Rogers and Marres, 2000, 2002). Studies of this type enable understanding of network structures on the web, but inferring the meaning or 'substance' of those network structures can be difficult to infer from large-scale mapping studies.

More recently, a third set of approaches to web analysis has emerged that takes hyperlink relationality into account in more nuanced ways. We refer to this set of approaches for analyzing multi-actor, cross-site action on the web as *sociocultural* analyses of the web (see several examples in Beaulieu and Park, 2003). Lindlof and Shatzer (1998) point in this direction in their article calling for new strategies of media ethnography in 'virtual space'. Hine (2000) presents a good example of sociocultural analysis of cross-site action on the web. Similarly, Howard's (2002) conceptualization of network ethnography reflects methodological sensitivity to processes of web production. By appropriating the term 'sociocultural' to describe this set of approaches, we seek to highlight the attention paid in this genre of web studies to the hyperlinked context(s) and situatedness of websites – and to the aims, strategies and identity-construction processes of website producers – as sites are produced, maintained and/or mediated through links.

WEB SPHERE ANALYSIS

Our own work has benefited from the methodological groundwork that has been established by our colleagues in web studies. We are developing a

multi-method approach called 'web sphere analysis' that enables analysis of communicative actions and relations between web producers and users developmentally over time. We conceptualize a web sphere as not simply a collection of websites, but as a hyperlinked set of dynamically-defined digital resources that span multiple websites and are deemed relevant, or related, to a central theme or 'object'. The boundaries of a web sphere are delimited by a shared object-orientation and a temporal framework. Web sphere analysis is an analytic strategy that includes relations between producers and users of web materials, as potentiated and mediated by the structural and feature elements of websites, hypertexts, and the links between them (Foot and Schneider, 2002; Foot et al., 2003).

The most crucial element in this definition of web sphere is the dynamic nature of the sites to be included. This dynamism comes from two sources. First, the researchers involved in identifying the boundaries of the sphere are likely to find continuously new sites to be included within it. Second, the notion of defining a web sphere is recursive, in that pages that are referenced by other included sites, as well as pages that reference included sites, are considered as part of the sphere under evaluation. Thus, as a web sphere is analyzed over time (ideally via an archive that enables retrospective analysis), its boundaries are dynamically shaped by both researchers' identification strategies and changes in the sites themselves.

The web sphere can function as a macro unit of analysis, by which historical and/or inter-sphere comparisons can be made. For example, the web sphere of the 2000 elections in the United States can be analyzed comparatively with the electoral web sphere of 2002 and those that develop in later years, as well as with electoral web spheres in other countries. Alternatively and/or simultaneously, other, more micro units such as texts, features and/or links can be employed in analyses within a web sphere (Schneider and Foot, 2002; Schneider and Foot, 2003).

Web sphere analysis is an analytic strategy that, when fully implemented, includes analysis of the relations between producers and users of web materials as potentiated and mediated by the structural and feature elements of websites, hypertexts, and the links between them. In a nutshell, the multi-method approach of web sphere analysis consists of the following elements. Websites related to the object or theme of the sphere are identified, captured in their hyperlinked context, and archived with some periodicity for contemporaneous and retrospective analyses. The archived sites are annotated with human and/or computer-generated 'notes' of various kinds, which creates a set of metadata. These metadata correspond to the unit(s) and level(s) of analysis anticipated by the researcher(s). Sorting and retrieval of the integrated metadata and URL files is accomplished through several computer-assisted techniques. Interviews of various kinds are conducted with the producers and users of the websites in the identified

sphere, to be triangulated with web media data in the interpretation of the sphere.

CONCLUSIONS

The emergence of the internet, and especially the web, has challenged scholars conducting research to both adapt familiar methods and develop innovative approaches that account for the unique aspects of the web. This uniqueness includes both the nature of the communicative processes that it engenders, and the challenges that are posed in order to create research repositories allowing robust analyses (that are representative and reproducible) to proceed.

Methodological innovations have emerged in correspondence with the properties of these new media applications. This analysis has highlighted some methodological trends. Earlier studies of the internet tended to focus either on users and/or usage patterns, or on media and production characteristics. Within the user studies genre, the predominant methods were various forms of textual or discourse analysis and participant observation, with online surveys and experiments emerging later. Overarching these trends is a shift toward methods that recognize the co-productive nature of new media – thus the duality of users and producers – and the potential for digital media productions to be simultaneously inscriptions of communicative action and structures for action, especially on the web.

The emergence of web archiving techniques that are designed to facilitate scholarly analysis integrates researchers into archiving activities. Traditionally, the work of archivists proceeded largely independently of the scholars who would be expected to use the archived materials as the basis of research work at a later time. Given the cost and complexity of web archiving, an alternative approach is emerging that attempts to integrate researchers into archiving activities. If scholars join with archivists to identify web objects of interest, and to delineate strategies for building archives that support scholarly activities, the basis of future research efforts is likely to be enhanced, and new methods for web studies developed.

References

- Arms, W., R. Adkins, C. Ammen and A. Hayes (2001) 'Collecting and Preserving the Web: the MINERVA Prototype', *RLG DigiNews*, 15 April, 5(2), URL (consulted August 2003): <http://www.rlg.org/preserv/diginews/diginews5-2.html>
- Baym, N. (1999) *Tune in, Log On: Soaps, Fandom and On-line Community*. Thousand Oaks, CA: Sage.
- Beaulieu, A. and H.W. Park (eds) (2003) 'Special Issue: Internet Networks: the Form and the Feel', *Journal of Computer-mediated Communication* 8(4), available online: <http://www.ascusc.org/jcmc/vol8/issue4>.
- Benoit, W.J. and P.J. Benoit (2000) 'The Virtual Campaign: Presidential Primary Websites in Campaign 2000', *American Communication Journal* 3(3), available online: <http://acjournal.org/holdings/vol3/Iss3/curtain.html#4>

- Berners-Lee, T. (2000) *Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web*. New York: HarperCollins.
- Burbules, N.C. and T.A. Callister (2000) *Watch IT: the Risks and Promises of Information Technologies for Education*. Boulder, CO: Westview Press.
- Crowston, K. and M. Williams (2000) 'Reproduced and Emergent Genres of Communication on the World Wide Web', *The Information Society* 16(3): 201–15.
- D'Alessio, D. (1997) 'Use of the World Wide Web in the 1996 US Election', *Electoral Studies* 16(4): 489–500.
- D'Alessio, D. (2000) 'Adoption of the World Wide Web by American Political Candidates, 1996–1998', *Journal of Broadcasting and Electronic Media* 44(4): 556–68.
- Foot, K.A. and S.M. Schneider (2002) 'Online Action in Campaign 2000: an Exploratory Analysis of the U.S. Political Web Sphere', *Journal of Broadcasting & Electronic Media* 46(2): 222–44.
- Foot, K.A., S.M. Schneider, M. Dougherty, M. Xenos and E. Larsen (2003) 'Analyzing Linking Practices: Candidate Sites in the 2002 U.S. Electoral Web Sphere', *Journal of Computer-mediated Communication* 8(4), available online: <http://www.ascusc.org/jcmc/vol8/issue4/foot.html>
- Gauntlett, D. (ed.) (2000) *Web.studies: Rewiring Media Studies for the Digital Age*. New York: Oxford University Press.
- Hakken, D. (1999) *Cyborgs@Cyberspace? An Ethnographer Looks Into the Future*. London: Routledge.
- Hansen, G. (2000) 'Internet Presidential Campaigning: the Influences of Candidate Internet Sites on the 2000 Elections', paper presented at the National Communication Association, Seattle, WA, November.
- Hecht, M.L., S.R. Corman and M. Miller-Rassulo (1993) 'An Evaluation of the Drug Resistance Project: a Comparison of Film Versus Live Performance Media', *Health Communication* 5(2): 75–88.
- Hine, C. (2000) *Virtual Ethnography*. Thousand Oaks, CA: Sage.
- Howard, P. (2002) 'Network Ethnography and Hypermedia Organization: New Organizations, New Media, New Myths', *New Media & Society* 4(4): 550–74.
- Howard, P.N. and S. Jones (eds) (2003) *Society Online: the Internet in Context*. Thousand Oaks, CA: Sage.
- Iyengar, S. (2002) 'Experimental Designs for Political Communication Research: from Shopping Malls to the Internet', paper presented at Workshop in Mass Media Economics, Department of Political Science, London School of Economics, 25–26 June, available at: <http://pcl.stanford.edu/common/docs/research/iyengar/2002/expdes2002.pdf>.
- Jackson, M. (1997) 'Assessing the Structure of the Communication on the World Wide Web', *Journal of Computer-mediated Communication* 3(1), available online: <http://www.ascusc.org/jcmc/vol3/issue1/jackson.html>
- Jones, S. (ed.) (1999) *Doing Internet Research: Critical Issues and Methods for Examining the Net*. Thousand Oaks, CA: Sage.
- Kahle, B. (1997) 'Preserving the Internet', *Scientific American* 276(3): 82–3.
- Lindlof, T.R. and M.J. Shatzer (1998) 'Media Ethnography in Virtual Space: Strategies, Limits, and Possibilities', *Journal of Broadcasting and Electronic Media* 42(2): 170–89.
- Lyman, P. (2002) 'Archiving the World Wide Web', *Building a National Strategy for Digital Preservation*, April, URL (consulted August 2003): <http://www.clir.org/pubs/reports/pub106/web.html>

- Lyman, P. and B. Kahle (1998) 'Archiving Digital Cultural Artifacts: Organizing an Agenda for Action', *D-Lib Magazine*, URL (consulted August 2003): <http://www.dlib.org/dlib/july98/07lyman.html>
- McMillan, S.J. (1999) 'Health Communication and the Internet: Relationships Between Interactive Characteristics of the Medium and Site Creators, Content, and Purpose', *Health Communication* 11(4): 375–90.
- Mann, C. and F. Stewart (2000) *Internet Communication and Qualitative Research Online: a Handbook for Researching Online*. London: Sage.
- Markham, A.N. (1998) *Life Online: Researching Real Experience in Virtual Space*. Walnut Creek, CA: Altamira Press.
- Mitra, A. (1999) 'Characteristics of the WWW Text: Tracing Discursive Strategies', *Journal of Computer-mediated Communication* 5(1), available online: <http://www.ascusc.org/jcmc/vol5/issue1/mitra.html>.
- Mitra, A. and E. Cohen (1999) 'Analyzing the Web: Directions and Challenges', in S. Jones (ed.) *Doing Internet Research: Critical Issues and Methods for Examining the Net*, pp. 179–202. Thousand Oaks, CA: Sage.
- Odlyzko, A. (2001) 'Content is Not King', *First Monday* 6(2), URL (consulted August 2003): http://firstmonday.org/issues/issue6_2/odlyzko/index.html.
- Park, H.W. (2003) 'What is Hyperlink Network Analysis? New Method for the Study of Social Structure on the Web', *Connections* 25(1): 49–62.
- Parks, M. and K. Floyd (1996) 'Making Friends in Cyberspace', *Journal of Computer-mediated Communication* 1(4), available online: <http://www.ascusc.org/jcmc/vol1/issue4/vol1no4.html>.
- Park, H.W. and M. Thelwall (2003) 'Hyperlink Analyses of the World Wide Web: a Review', *Journal of Computer-Mediated Communication* 8(4), available online: <http://www.ascusc.org/jcmc/vol8/issue4/park.html>
- Price, V. and J.N. Capella (2001) 'Online Deliberation and its Influence: the Electronic Dialogue Project in Campaign 2000', paper presented at the American Association for Public Opinion Research, May, Canada.
- Rogers, R. and N. Marres (2000) 'Landscaping Climate Change: a Mapping Technique for Understanding Science and Technology Debates on the World Wide Web', *Public Understanding of Science* 9(2): 141–63.
- Rogers, R. and N. Marres (2002) 'French Scandals on the Web, and on the Streets: a Small Experiment in Stretching the Limits of Reported Reality', *Asian Journal of Social Science* 30(2): 339–53.
- Schmidt, W.C. (1997) 'World Wide Web Survey Research: Benefits, Potential Problems, and Solutions', *Behavior Research Methods, Instruments & Computers* 29(2): 274–9.
- Schneider, S. M. and K.A. Foot (2002) 'Online Structure for Political Action: Exploring Presidential Web Sites from the 2000 American Election', *Javnost (The Public)* 9(2): 43–60.
- Schneider, S.M. and K.A. Foot (2003) 'Crisis Communication and New Media: the Web After September 11', in P.N. Howard and S. Jones (eds) *Society Online: the Internet in Context*, pp. 137–54. Thousand Oaks, CA: Sage.
- Sillaman, L. (2000) 'The Digital Campaign Trail: Candidate Images on Campaign Websites', master's thesis, Annenberg School for Communication, University of Pennsylvania.
- Smith, C. (1997) 'Casting the Net: Surveying an Internet Population', *Journal of Computer-mediated Communication* 3(1), available online: <http://www.ascusc.org/jcmc/vol3/issue1/smith.html>

- Stowkowski, P.A. (2002) 'Languages of Place and Discourses of Power: Constructing New Senses of Place', *Journal of Leisure Research* 34(4): 368–82.
- Stromer-Galley, J. and K.A. Foot (2002) 'Citizen Perceptions of Online Interactivity and Implications for Political Campaign Communication', *Journal of Computer-mediated Communication* 8(1), available online: <http://www.ascusc.org/jcmc/vol8/issue1/stromerandfoot.html>
- Warnick, B. (1998) 'Appearance or Reality? Political Parody on the Web in Campaign '96', *Critical Studies on Mass Communication* 15(3): 306–24.
- Warnick, B. (2001) *Critical Literacy in a Digital Era: Technology, Rhetoric, and the Public Sphere*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Witte, J., L. Amoroso and P.N. Howard (2000) 'Method and Representation in Internet-based Survey Tools: Mobility, Community, and Cultural Identity in Survey 2000', *Social Science Computer Review* 18(2): 179–195.
- Yun, G.W. and C. Trumbo (2000) 'Comparative Response to a Survey Executed by Post, E-mail and Web Form', *Journal of Computer-mediated Communication* 6(1), available online: <http://www.ascusc.org/jcmc/vol6/issue1/yun.html>
-

STEVEN M. SCHNEIDER is a visiting associate professor in the Department of Language, Literature and Communication at Rensselaer (Troy, NY) for the 2003–2004 year, and is Associate Professor at SUNY Institute of Technology. He is currently co-editor of politicalweb.info, and co-director of WebArchivist.org. Schneider has been studying the relationship between political life and the internet since 1988, and received a PhD in Political Science from Massachusetts Institute of Technology.

Address: SUNY Institute of Technology, Utica, NY 13504, USA. [email: steve@sunyit.edu]

KIRSTEN FOOT is an assistant professor in the Department of Communication at the University of Washington. As co-director of the WebArchivist.org research group, she is developing new techniques for studying social and political action on the web. Her research interests include co-production and mobilization on the web, and online campaigning practices. She co-edits the *Acting With Technology* series at MIT Press, and her work has been published in *Communication Theory*, *Journal of Broadcasting and Electronic Media*, and *Journal of Computer-Mediated Communication*.

Address: Department of Communication, Box 353740, University of Washington, Seattle, WA 98195, USA. [email: kfoot@u.washington.edu]

M. Thelwall, L. Vaughan and L. Björneborn (2005). "Webometrics," in B. Cronin (ed.), *Annual Review of Information Science and Technology*, 39, Medford, NJ: Information today, 81-135.

Webometrics

Mike Thelwall
University of Wolverhampton

Liwen Vaughan
University of Western Ontario

Lennart Björneborn
Royal School of Library and Information Science

Introduction

Webometrics, the quantitative study of Web-related phenomena, emerged from the realization that methods originally designed for bibliometric analysis of scientific journal article citation patterns could be applied to the Web, with commercial search engines providing the raw data. Almind and Ingwersen (1997) defined the field and gave it its name. Other pioneers included Rodríguez Gairín (1997) and Aguillo (1998). Larson (1996) undertook exploratory link structure analysis, as did Rousseau (1997). Webometrics encompasses research from fields beyond information science such as communication studies, statistical physics, and computer science. In this review we concentrate on link analysis, but also cover other aspects of webometrics, including Web log file analysis.

One theme that runs through this chapter is the messiness of Web data and the need for data cleansing heuristics. The uncontrolled Web creates numerous problems in the interpretation of results, for instance, from the automatic creation or replication of links. The loose connection between top-level domain specifications (e.g., .com, .edu, and .org) and their actual content is also a frustrating problem. For example, many .com sites contain noncommercial content, although .com is ostensibly the main commercial top-level domain. Indeed, a skeptical researcher could claim that obstacles of this kind are so great that all Web analyses lack value. As will be seen, one response to this view, a view shared by critics of evaluative bibliometrics, is to demonstrate that Web data correlate significantly with some non-Web data in order to prove that the Web data are not wholly random. A practical response has been to develop increasingly sophisticated data cleansing techniques and multiple data analysis methods.

This review is split into four parts: basic concepts and methods, scholarly communication on the Web, general and commercial Web use, and topological modeling and mining of the Web. Webometrics is a new field based on analysis of new forms of data; therefore, methods of data collection and processing have been prominent in many studies. The second part, scholarly communication on the Web, is concerned with using link analysis to identify patterns in academic or scholarly Web spaces. Almost all of these studies have direct analogies in traditional bibliometrics and have sought to develop effective methods and validation techniques, the latter being an issue of particular concern on the Web. The general and commercial Web use section reviews link analysis studies that have used techniques similar to those applied to academic Web spaces. Some have their origins in social network analysis rather than information science, producing an interesting complementary perspective. The section also includes quantitative studies of the size of the “whole” Web and Web server log analysis. The final section, topological modeling and mining of the Web, covers mathematical approaches to modeling the growth of the Web or its internal link structure, mostly the product of computer science and statistical physics research. It concludes with detailed interpretations of “small-world” linking phenomena, an information science contribution to topological modeling of the Web.

Webometrics, Bibliometrics, and Informetrics

Being a global document network initially developed for scholarly use (Berners-Lee & Cailliau, 1990), the Web constitutes an obvious research topic for bibliometrics, scientometrics, and informetrics. A range of new terms for the emerging research field have been proposed since the mid-1990s, for instance, *netometrics* (Bossy, 1995), *webometry* (Abraham, 1996), *internetometrics* (Almind & Ingwersen, 1996), *webometrics* (Almind & Ingwersen, 1997), *cybermetrics* (also the name of the journal started in 1997 by Isidro Aguillo), *Web bibliometry* (Chakrabarti, Joshi, Punera, & Pennock, 2002), and *Web metrics* (the term used in computer science, for example by Dhyani, Keong & Bhowmick, 2002). Webometrics and cybermetrics are currently the two most widely adopted terms in information science, often used as synonyms.

Björneborn and Ingwersen (in press) have proposed a differentiated terminology, distinguishing between studies of the Web and studies of *all* Internet applications. They used an information science-related definition of webometrics as “the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the WWW drawing on bibliometric and informetric approaches” (Björneborn & Ingwersen, in press). This definition thus covers quantitative aspects of both the construction and usage sides of the Web, embracing the four main areas of webometric research: 1) Web page content analysis, 2) Web link structure analysis, 3) Web usage analysis (e.g., exploiting log files of users’ searching and browsing behavior), and

4) Web technology analysis (including search engine performance). The definition encompasses hybrid forms. For example, Pirolli, Pitkow, and Rao (1996) explored Web analysis techniques for automatic categorization utilizing link graph topology, text content, and metadata similarity, as well as usage data. All four main research areas include longitudinal studies of changes on the dynamic Web, for example, of page contents, link structures, and usage patterns. So-called *Web archaeology* (Björneborn & Ingwersen, 2001) could be important for recovering historical Web developments, for instance, by means of the Internet Archive (<http://www.archive.org>), an approach already used in webometrics (Björneborn, 2003; Thelwall & Vaughan, 2004; Vaughan & Thelwall, 2003).

Furthermore, Björneborn and Ingwersen (in press) have proposed cybermetrics as a generic term for “the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the whole Internet, drawing on bibliometric and informetric approaches.” Cybermetrics thus encompasses statistical studies of discussion groups, mailing lists, and other computer-mediated communication on the Internet (e.g., Bar-Ilan, 1997; Hernández-Borges, Pareras, & Jiménez, 1997; S. C. Herring, 2002; Matzat, 1998) including the Web. In addition to covering all computer-mediated communication using Internet applications, this definition of cybermetrics also covers quantitative measures of the Internet backbone technology, topology, and traffic (Molyneux & Williams, 1999). The breadth of coverage of cybermetrics and webometrics implies considerable overlap with proliferating computer science-based approaches in analyses of Web contents, link structures, Web usage, and Web technologies. A range of such approaches has emerged since the mid-1990s with names like *Cyber geography / Cyber cartography* (e.g., Dodge, 1999; Dodge & Kitchin, 2001; Girardin, 1996), *Web ecology* (e.g., Chi, Pitkow, Mackinlay, Pirolli, Gossweiler, & Card, 1998; Huberman, 2001), *Web mining* (e.g., Chen & Chau, 2004; Etzioni, 1996; Kosala & Blockeel, 2000), *Web graph analysis* (e.g., Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, et al., 2000; Chakrabarti, Dom, Kumar, Raghavan, Rajagopalan, Tomkins, et al., 1999; Kleinberg, Kumar, Raghavan, Rajagopalan, & Tomkins, 1999), and *Web intelligence* (e.g., Yao, Zhong, Liu, & Ohsuga, 2001). One reason for using the term *webometrics* is to show the connection with bibliometrics and informetrics and to stress an information science perspective.

There are different conceptions of informetrics, bibliometrics, and scientometrics. The diagram in Figure 3.1 (Björneborn & Ingwersen, in press) shows the field of *informetrics* embracing the overlapping fields of *bibliometrics* and *scientometrics* following widely adopted definitions by, among others, Brookes (1990), Egghe and Rousseau (1990), and Tague-Sutcliffe (1992). According to Tague-Sutcliffe (1992, p. 1), informetrics is “the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists.” Bibliometrics is defined as “the study of the quantitative aspects

of the production, dissemination and use of recorded information” and scientometrics as “the study of the quantitative aspects of science as a discipline or economic activity” (Tague-Sutcliffe, 1992, p. 1). In the figure, politico-economic aspects of scientometrics are covered by the part of the scientometric ellipse lying outside the bibliometric one.

Webometrics may be seen as entirely encompassed by bibliometrics, because Web documents, whether text or multimedia, are recorded information (cf. Tague-Sutcliffe’s definition of bibliometrics) stored on Web servers. This recording may be temporary only, just as not all paper documents are properly archived. In the diagram, webometrics is partially covered by scientometrics, because many scholarly activities today are Web-based. Furthermore, webometrics is contained within the field of cybermetrics as defined here.

In Figure 3.1, the field of cybermetrics exceeds the boundaries of bibliometrics because some activities in cyberspace are not normally recorded, but communicated synchronously, as in chat rooms. Cybermetric studies of such activities still fit in the generic field of informetrics as the study of the quantitative aspects of information “in any form” and “in any social group,” as Tague-Sutcliffe (1992) stated.

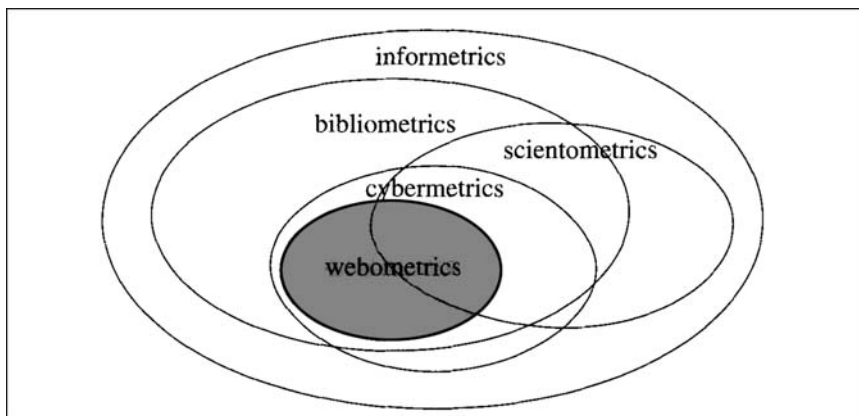


Figure 3.1 Infor-, biblio-, sciento-, cyber-, and webometrics (Björneborn & Ingwersen, in press). The sizes of the overlapping ellipses are made for sake of clarity only.

The inclusion of webometrics expands the field of bibliometrics because webometrics will inevitably contribute further methodological developments. As ideas rooted in bibliometrics, scientometrics, and informetrics have contributed to the emergence of webometrics, insights from webometrics will likely contribute to the development of these more established fields.

Scope of Coverage and Related Reviews

Although webometrics is still a young field, it has gained recognition: The Association for Library and Information Science Education lists webometrics in its Research Areas Classification and the *Encyclopedia of Social Measurement* includes an article on the topic (Vaughan, in press). Already many review articles offer partial coverage, including two *ARIST* chapters. The first is by Molyneux and Williams (1999), "Measuring the Internet." So much has changed since that chapter was written that it is now dated. The bibliometrics chapter by Borgman and Furner (2002) explicitly cast Web links as structurally equivalent to journal citations, developing a terminology in which both were referred to as links. Web research within and beyond information science was covered and there will be many connections from that chapter to this—especially the scholarly communication section—despite a relatively small overlap in articles reviewed. This similarity is important because many issues from standard bibliometrics are mirrored in Web behaviors and phenomena. A special issue of the *Journal of the American Society for Information Science and Technology* on webometrics is set to appear in 2004. It will contain a wide variety of different approaches to the quantitative study of the Web, some of which are reviewed in this chapter.

Bar-Ilan and Peritz (2002, p. 371) have published a review of "informetric theories and methods for exploring the Internet," which is similar in scope to ours but includes non-Web, Internet research, is more focused on general informetric techniques, and does not reflect recent significant advances in hyperlink analysis. Web metrics are surveyed from a computer science perspective by Dhyani, Keong, and Bhowmick (2002), covering mathematical techniques for measuring a wide variety of online phenomena. Rasmussen's (2003) *ARIST* chapter, "Indexing and Retrieval for the Web," provides useful additional information on Web crawlers and search engines. Our review differs from these in another regard: Its primary focus is on different types of Web link analysis.

Several general webometrics articles were published in 2003, reflecting a widening recognition of the importance of the topic and the existence of a critical mass of research. Park and Thelwall (2003) compared information science approaches to studying the Web to those from social network analysis. They found that information scientists emphasized data validation and the study of methodological issues, whereas social network analysts experimented with transferring existing theory to the Web. Wilkinson, Thelwall, and Li (2003) introduced Web-based quantitative methods to social scientists for use in their Web research. This coincided with the release of a special tool for social science link structure analysis, SocSciBot (<http://cybermetrics.wlv.ac.uk/socscibot>) and included a review section as well as a more prescriptive "how to do it" conclusion. Ingwersen's (1998) Web Impact Factor calculation was singled out for detailed coverage by Li (2003). Peres Vanti (2002) presented a general review.

Other review articles cover closely related topics. Search engine research is covered by Bar-Ilan (2004), search engines themselves by Arasu, Cho, Garcia-Molina, Paepcke, and Raghavan (2001), and data collection techniques in general by Bar-Ilan (2001) and Thelwall (2002d). Henzinger (2001) reviewed link structure analysis from a computer science perspective, showing how links could be used in search engine ranking algorithms. Barabási (2002) and Huberman (2001) have written popular science books explaining current research into mathematical modeling of the growth of the Web. Web mining (Chen & Chau, 2004; Kosala & Blockeel, 2000) and Web intelligence (Yao et al., 2001) are also relevant.

Online communication studies can contribute to webometric research by providing useful context information and explanatory theories. S. C. Herring (2002) gave an overview of computer-mediated communication on the Internet, taking a social science perspective and focusing mostly on non-Web media such as e-mail and newsgroups. One of the key general findings was that the use of new technology was very context specific, determined by users' needs rather than the technology. Ellis, Oldridge, and Vasconcelos (2004) explored communication in electronic communities. Finholt (2002) reviewed electronic virtual laboratories for scientists to share equipment or data (collaboratories). Finally, Kling and Callahan (2003) discussed e-journals and scholarly communication, which was useful background for our e-journals subsection.

Our review not only covers more recent research (including some currently unpublished research) but also has two areas of special emphasis: link analysis in academic Web spaces and basic concepts and methods. Other areas are covered, but less comprehensively, and either draw from or feed into the two main themes.

Basic Concepts and Methods

This section contains much terminology and many technical details. It is intended to provide a coherent basis for future webometric research and background to the studies reported. Some readers may wish to skip to the next section, "Scholarly Communication on the Web" on a first reading and return to this one later.

Terminology

Basic Link Terminology

Many new terms are used in the webometrics literature. For instance, a link received by a Web node (the network term "node" here covers a unit of analysis, such as a Web page, directory, Web site, or an entire top level domain of a country) has been named, variously, *incoming link*, *inbound link*, *inward link*, *back link*, and *sitation*; the latter term (McKiernan, 1996; Rousseau, 1997) clearly derives from bibliometric/citation analysis. An example of more problematic terminology is the two opposing meanings

of an *external link*: either as a link pointing out of a Web site or a link pointing into a site. We recommend the consistent basic webometric terminology of Björneborn and Ingwersen (in press) for link relations between Web nodes, as briefly outlined in Figure 3.2. The proposed terminology has origins in graph theory, social network analysis, and bibliometrics.

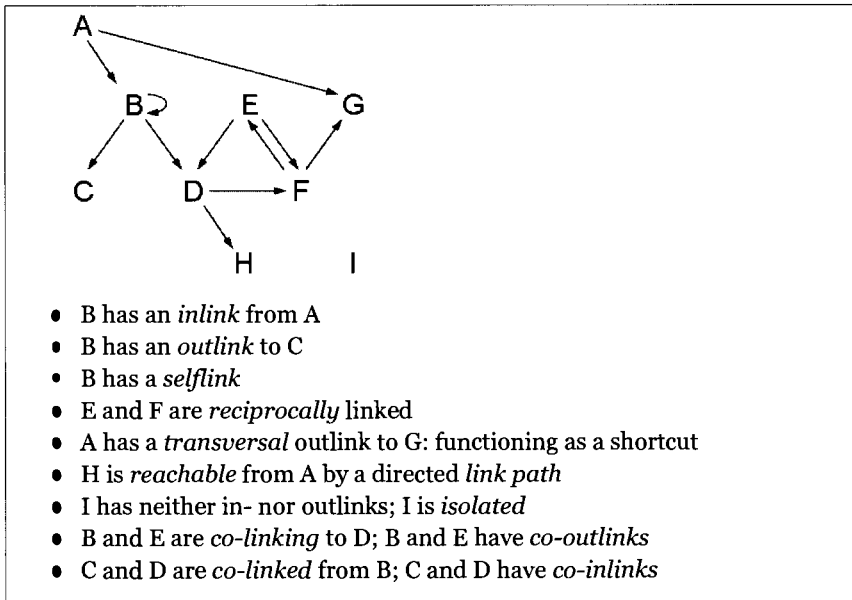


Figure 3.2

Basic webometric link terminology (see Björneborn & Ingwersen [in press] for a more detailed legend). The letters may represent different Web node levels, for example, Web pages, Web directories, Web sites, or top level domains of countries or generic sectors.

The terms *outlink* and *inlink* are commonly used in computer science-based Web studies (e.g., Broder et al., 2000; Chen, Newman, Newman, & Rada, 1998; Pirolli et al., 1996). The term *outlink* implies that a directed link and its two adjacent nodes are viewed from the source node providing the link, analogous with the use of the term *reference* in bibliometrics. A corresponding analogy exists between the terms *inlink* and *citation*, with the target node as the spectator's viewpoint. In the conceptual framework of Björneborn and Ingwersen (in press), a link crossing a Web site border, like link *e* in Figure 3.3, is called a *site outlink* or a *site inlink* depending on the perspective of the spectator.

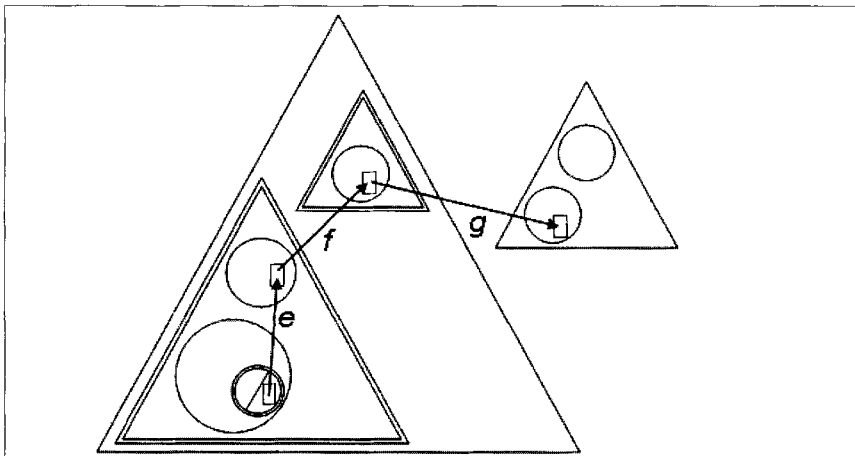


Figure 3.3 Simplified Web node diagram illustrating basic node levels (Björneborn & Ingwersen, in press).

On the Web, *selflinks* are used for a wider range of purposes than self-citations in scientific literature. Page selflinks point from one section to another within the same page. Site selflinks (also known as *internal links*) are typically navigational pointers from one page to another within the same site. Most links on the Web connect Web pages containing cognate topics (Davison, 2000). However, some links may break a typical linkage pattern in a Web node neighborhood and connect dissimilar topical domains. Such (loosely defined) *transversal links* (Björneborn, 2001, 2003) function as cross-topic shortcuts and may affect so-called small-world phenomena on the Web (cf. the section on small worlds).

The two *co-linked* Web nodes C and D in Figure 3.2 with *co-inlinks* from the same source node are analogous to the bibliometric concept of *co-citation* (Small, 1973). Correspondingly, the two *co-linking* nodes B and E having *co-outlinks* to the same target node are analogous to *bibliographic coupling* (Kessler, 1963). *Co-links* is proposed as a generic term covering both concepts (Björneborn & Ingwersen, in press)

Basic Web Node Terminology and Diagrams

In webometric studies it may be useful to visualize relations between different units of analysis. Figure 3.2 illustrates some basic building blocks in a consistent Web node framework (Björneborn & Ingwersen, in press). In the diagram, four basic Web node levels are denoted with simple geometrical figures: *quadrangles* (Web pages), *diagonal lines* (Web directories), *circles* (Web sites) and *triangles* (country or generic top level domains, TLDs). Sublevels within each of the four basic node levels are

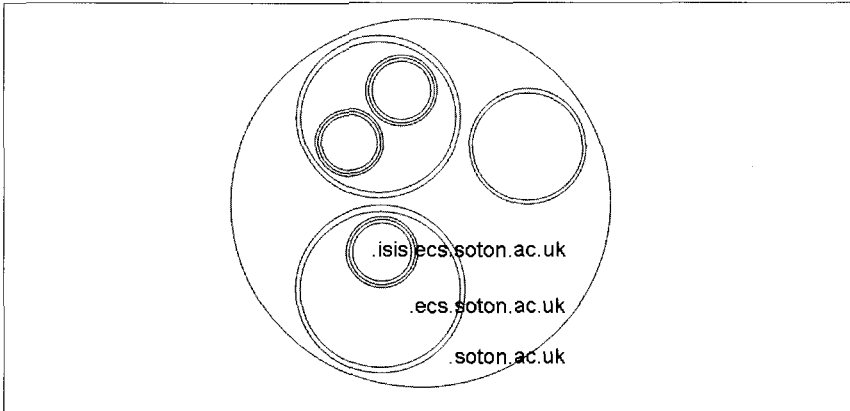


Figure 3.4 Simplified Web node diagram of a Web site containing subsites and sub-subsites (Bjørneborn & Ingwersen, in press).

denoted with additional borderlines in the corresponding geometrical figure. For example, a triangle with a double borderline denotes a generic second level domain (SLD), also known as a sub-TLD, assigned by many countries to educational, commercial, governmental, and other sectors, for instance: *.ac.uk*, *.co.uk*, *.ac.jp*, *.edu.au*.

The simplistic Web node diagram in Figure 3.2 shows a page *P* located in a directory of a subsite in a sub-TLD. The page has a site outlink, *e*, to a page at a site in the same sub-TLD. The outlinked page in turn is outlinking, *f*, to a page at a site in another sub-TLD in the same country. The link path *e-f-g* ends at a page at a site in another TLD.

Zooming in on a single Web site may reveal several subunits in the shape of subsites, sub-subsites, and so on, as indicated by hierarchically derivative domain names. For instance, as shown in Figure 3.4, the sub-subsite of The Image, Speech and Intelligent Systems Research Group (*isis.ecs.soton.ac.uk*) is located within the Department of Electronics and Computer Science (*ecs.soton.ac.uk*), one of the subsites at the University of Southampton, U.K. (*soton.ac.uk*). Subsites and sub-subsites are denoted as circles with double and triple borderlines, respectively. Subordinate sublevels would logically be denoted with additional number of borderlines. For the sake of simplicity, the diagram does not reflect actual numbers and sizes of elements.

Although some Web sites subdivide into derivative domain names, as noted, others locate the same type of subunits in folder directories. Obviously, such diverse allocation and naming practices complicate comparability in webometric studies. In Figure 3.5 directories, subdirectories, and so on, are denoted by one or more diagonal lines resembling URL slashes and reflecting the number of directory levels below the URL root level.

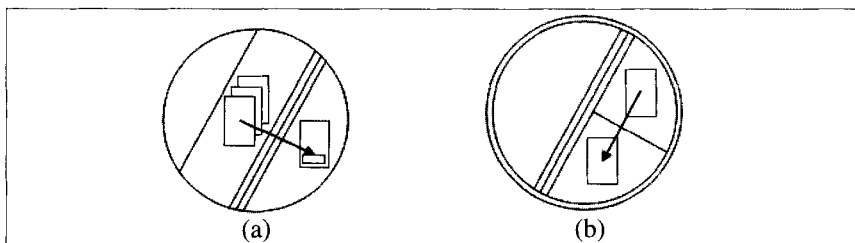


Figure 3.5 Simplified Web node diagrams of a Web site and a subsite with links between different directory levels including page subelements (Bjørneborn & Ingwersen, in press).

Web pages may also consist of subelements, such as text sections, and frames. Additional bands illustrate such page subelements as in the targets of the page selflink h and the page outlink i from the two sibling Web pages in the same directory in Figure 3.5a. More complex and numerous linkages within a site or subsite can be illustrated by combinations of elements in Figure 3.5, showing links between pages located either at different directory levels (Figure 3.5a) or in sibling directories at the same level (Figure 3.5b) in the Web site file hierarchies.

Naturally, any diagrammatic representation of large-scale hypertext structures will become too tangled to be of practical use, even less to be interpreted in any quantitative way. However, the proposed Web node diagrams with their simple and intuitive geometrical figures are intended to be used to emphasize and illustrate qualitative differences between investigated Web node levels. Moreover, the diagrams can illustrate important structural aspects of limited subgraphs of a given Web space.

Units of Analysis

The Definition of the Web

In this subsection we review different units of analysis in webometrics, after discussing the problematic issue of a precise definition of the Web itself. The Web is a loose concept (Berners-Lee, 1999) for which many definitions are possible (e.g., Boutell, 2003; Lavoie & Nielsen, 1999; webopedia.com, 2003) and the trend seems to be toward very inclusive definitions, including anything that can be accessed through a Web browser. It is important to be aware of this problem when devising a study and evaluating others' results. For example, non-HTML, automatically generated pages such as for search engine results, and password-protected pages could be reasonably included or excluded. Researchers should clearly state which definition they are using; but we do not recommend any one definition, only an awareness of the potential importance of differences. If the most generous definition were chosen, the size of the Web would be practically infinite due to the inclusion of

dynamically created pages, pages that are created by Web servers in response to queries. The terms “Invisible Web” and “Deep Web” have been coined for the vast quantity of Web pages that are accessible with a Web browser but are difficult to find because they are of a type not indexed by search engines (Bergman, 2000). Invisible pages are a particular problem for large-scale studies of the Web, but researchers can explicitly state that they are excluded by a precise definition of the area studied. As an example of this, Lawrence and Giles (1999a) define their scope as pages on sites that can be reached by following links from the site home page. In practice, something similar to this often defines the scope of webometric studies.

Alternative Document Models

When counting objects on the Web, a decision needs to be made about the most appropriate unit of measure. Often this will be either the Web page or Web site: How many business Web sites are there in the U.S.? How many Web pages are there in the Stanford University Web site? But both of these have their problems and alternatives are possible.

There are technical difficulties in identifying Web pages, some associated with the definition of the Web itself, some to do with HTML and others concerned with a broader interpretation of what a document should be. An HTML-related problem is that a single page may be made up of multiple files if the frameset HTML tag is used. Counting files irrespective of whether they are in a frameset is a possible solution, but since individual frames in a frameset can change independently, there is no perfect answer (Thelwall, 2002d). A conceptual problem is that a Web author can choose various ways to divide a work. An online book, for example, could be presented as one huge Web page or thousands of section-based pages. An ideal solution would perhaps be to count pages or collections of pages by genre, so that in both of these cases what would be counted would be one instance of the online book genre. But genre identification on the Web is a highly complex, unsolved problem (Crowston & Williams, 2000; Rehm, 2002) and so heuristics have been developed as a crude substitute, notably the Alternative Document Models (ADMs) (Thelwall, 2002b). There are four main ADMs in use, the page, directory, domain, and site models.

The page ADM is the default unit, based upon individual files. With the directory ADM, all pages in the same (virtual) directory identified through the URL file name path are counted as one unit. In the online book example, if the Web page sections were stored in a single directory, then a thousand-page online book would be counted as a single entity, a Web directory. Clearly this is a heuristic, given that one might also store an entire library in a single directory. The domain ADM (similar to a subsite in Figure 3.4) aggregates all pages with the same domain name in their URL into a common entity. The site ADM (originally called the university ADM) allows multiple domain names in the same document by specifying only the domain name ending. The site ADM of the

University of Southampton in Figure 3.4 would include all pages from subsites, sub-subsites, and so on, with URLs that contain domain names ending in .soton.ac.uk.

Another unit of analysis has been proposed and extensively used by Aguillo (1998), the *seat* (named subsequently to its description in his article). This is similar to the site ADM and conventional understandings of the phrase “Web site” but contains elements of genre. A seat is a collection of Web pages created by a single person or organization for a specific purpose. Typically, seats will be larger than a single page and may be a collection of pages associated with an individual, however hosted on the Web. Seats will sometimes be the same as directory, domain, and site ADMs, but are more flexible. As a consequence, the concept is both more powerful and harder to apply automatically in practice.

When counting documents or links, the ADMs give the option of counting at different aggregation levels. For instance, the number of domains in a university site could be counted, as could the number of directories that contain a link to a given Web site. This additional flexibility can be used to bypass anomalies in the data and conceptual problems with counting pages, as reported in the subsections on university and departmental studies. The choice of ADM for any particular study depends upon the patterns of use of the Web in the area covered. Two techniques have been proposed. The first is to compare correlations between summary statistics calculated by different ADMs and an external data source (Thelwall & Harries, 2003; Thelwall & Wilkinson, 2003b). The second is a theoretical approach using assumptions about the process that causes document creation and introducing tests to assess the extent to which the assumptions fit the data (Thelwall, 2004, in press b).

Data Collection Methods

Personal Web Crawlers

When studying a small Web site, the researcher may be able to find all pages by browsing alone. On larger sites, an alternative is to use a commercial search engine or a Web crawler. A crawler, spider, or robot is software that can automatically and iteratively download pages and extract and download their links (Arasu et al., 2001; Heydon & Najork, 1999; Thelwall, 2001e). For example, a crawler could be given the URL of a site home page and fetch all pages on the site that can be found by following links from the home page. Crawler coverage will typically be less than 100 percent of a site, however. Crawlers find new pages by following links and, so, are likely to miss many isolated pages and collections of pages. Pages that are linked to can also be missed if the links are in a form that the crawler does not understand (for instance, JavaScript), if the pages are password protected, if the server is temporarily down, or if the owner requests that the page not be crawled (e.g., through the Robots Exclusion Protocol). In a large university Web site, the pages created by individual faculty or students would be missed unless they were integrated into the

site by devices such as server-generated online lists of faculty/student pages, or links manually added to official pages, perhaps by Web administrators. Crawlers run by commercial search engines can partially offset this problem by adding to a crawl the URLs of all pages previously fetched from the site, URLs obtained by links from other sites, and any URLs submitted to the search engine by users (Thelwall, 2001b). All crawlers also have to employ heuristics such as a maximum crawl depth, maximum path depth, or maximum page fetches per site to avoid infinite loops. Commercial search engines may also have their crawling policies influenced by economic factors, with the most extreme example being pay-for-inclusion schemes. Cothey (in press) has shown how different crawling policies can give very different results on the same site, emphasizing the importance of full reporting of crawling parameters research. He also differentiated between content crawling (to find all unique documents) and link crawling (to map the link relationships between pages, irrespective of duplication), showing how the scope of a crawl will have to be tailored to its purpose.

It is possible to guarantee coverage of all (static) pages in a large Web site if privileged access is gained to the computer on which the server is operating. This is possible using a non-Web program that would not work by following links but instead by trawling all of the directories that were within the scope of the server. This approach would be difficult to apply to studies of more than one Web site due to problems in obtaining permission for special access to others' computers.

Commercial Search Engines

Commercial search engines have been used extensively for webometric research and are particularly useful for counts of pages over large areas of the Web. Some, including AltaVista, allow advanced Boolean queries to combine information requests about pages, links, and domains (Almind & Ingwersen, 1997), which can be used to return counts of pages or link pages that satisfy given criteria. This is the facility that originally spawned webometrics (advanced search help pages of search engines should be consulted for details of the full range of queries available for any particular engine).

Commercial search engines have a number of drawbacks. They do not cover all of the Web or even the entire publicly indexable Web; coverage was under 17 percent for the largest search engine in 1999 (Lawrence & Giles, 1999a). Their crawling and reporting algorithms are commercial secrets, which is a problem for their use in scientific research. Search engine results have been found to be unreliable and to fluctuate, even over short periods of time (Bar-Ilan, 1999; Ingwersen, 1998; Mettrop & Nieuwenhuysen, 2001; Rousseau, 1999; Snyder & Rosenbaum, 1999). AltaVista has been found to be more reliable recently (Thelwall, 2001c; Vaughan & Thelwall, 2003) and to give good coverage of academic Web sites (Thelwall, 2001b). However, researchers should always take steps to assess the accuracy and coverage of results because the algorithms are

subject to change without notice and the effects of the advanced commands are not always fully documented. Accuracy can be checked by testing a sample of the pages returned to ensure that they do contain the features specified in the query. Coverage is more difficult to check, but for any given study it may be possible to use AltaVista's "host:" command to check the number of pages indexed by AltaVista in a set of key sites. Rousseau (1999) has proposed statistical techniques to compensate for fluctuations. Björneborn and Ingwersen (2001) recommend the use of sampling and filtering to clean up data from search engines, acknowledging that this is likely to be difficult. An alternative strategy is to employ multiple search engines and average the results; but this will increase theoretical coverage at the expense of weighting more highly the pages indexed by multiple engines (since they will be counted more than once) and making more time-consuming the process of checking the operating methods, accuracy, and coverage of the engines used. Alternatively, different search engines can be used in parallel, comparing the results of one against the other to see if internal differences substantially influence the results. This is an approach that deserves to be taken seriously. Of course, similarity in results does not rule out the possibility that both engines have similar flaws, perhaps for underlying technical reasons.

A future issue for the validity of data from search engines may be evolving commercial models, especially if they include pay-for-inclusion and pay-for-placement. There does not seem to be any logical reason why commercialization should not extend to universities and research groups paying for improved positioning in major engines. Clearly this would create problems for those using data from search engines to evaluate online impact. Additionally, the continued provision of advanced search interfaces should not be taken as a given: They may be discontinued, if judged not to be profitable. Both of these considerations are incentives for researchers to develop data collection techniques that do not rely upon AltaVista or any other search engine.

Commercial search engines display national biases in site coverage. One study of four countries showed that U.S. sites were much better covered by three major search engines than China, Taiwan, and Singapore (Vaughan & Thelwall, *in press b*). Site language did not seem to be a factor; lower coverage was related to the smaller number of indexed links targeting sites in countries that adopted the Web more recently, something research designs may need to take into account. A similar pattern was found for the Internet Archive (<http://www.archive.org>; Thelwall & Vaughan, 2004).

Time series data can be obtained from some search engines and also the Internet Archive. For example, AltaVista allows searches for pages by their last modified date (e.g., Leydesdorff & Curran, 2000). The results are only partially accurate because the search engine overwrites older versions when a page is modified; it thus returns only the date of the last modification. The Internet Archive Wayback Machine gives genuine time series data because it records all the dates on which a page

was found, keeping copies of every indexed version so that changes can be tracked over time (e.g., Vaughan & Thelwall, 2003). The Archive can be used for an earliest known date for any page but not for an accurate creation date. This is because the earliest recorded date is merely when it was first found by the Archive crawler, not when it was actually created. Björneborn (2003) obtained indicative ages of different topological components (cf. the Overall Web Topology section) of the U.K. academic Web space based on first indexing dates from the Archive. A disadvantage of the Wayback Machine, at the time of writing, is that it does not allow global Boolean queries beyond simple wildcard matches, unlike AltaVista. An attractive alternative is to employ a crawler to collect data over time for a preset list of Web pages or sites (e.g., Koehler, 2002), but this is not possible for retrospective studies.

Web Log Files

In addition to studying the Web itself, it can also be useful to collect data on how humans interact with it: on how it is used. Data collected could include behavioral studies of humans using the Web, although this is beyond the scope of our review. Another source of information is Web server log files. These record the requests sent to a Web server by users' browsers (and Web crawlers) and can be mined for information about how they are using the site. Typical applications of server log analysis would be to see which pages are most frequently viewed or to identify patterns of surfing with a view to improving site navigation. The principal drawback from a webometric perspective is that log files typically cover one site or all sites owned by the same server. However, access is often restricted, so these are generally not a good data source for studying Web site use.

Search engine log files are a special case because these can furnish information about what a user was interested in and where he went to find it. These files are also normally secret, but useful studies have been conducted by researchers who have been given access to sampled, individual log files (e.g., Silverstein, Henzinger, Marais, & Moricz, 1999; Spink, Ozmutlu, Ozmutlu, & Jansen, 2002; Spink, Wolfram, Jansen, & Saracevic, 2001).

There are many inherent difficulties in analyzing Web log files. The main problems are in identifying a user and determining a search session. In traditional information retrieval (IR) systems, users have to log in and out, so a user and a search session are explicitly identified. The Web is "stateless," which makes this identification extremely difficult. The main measures that have been used to determine a unique Web user are cookies (small quantities of information stored by the user's machine) and Internet Protocol (IP) addresses. Silverstein et al. (1999) used the former technique to study AltaVista users. For queries in which the user disabled the cookies, they used the pair "domain IP/Web browser used" as a substitute for the cookie. They acknowledge that the cookie method is imperfect because different people can use the same browser, for example, a public computer station in a library. Following a

user's search behavior over time is even more difficult because a person can use different computers, for instance, one at work and one at home. User identification by IP address can be even less precise because many users can share a single IP address in large Internet service providers and some organizations assign floating IP addresses to computers. More encouraging is the fact that 96 percent of the queries in Silverstein et al.'s (1999) study had cookie information. They analyzed their data in two ways, one for the full data set and the other for those with cookie information. The two results were very similar.

A common method for determining a search session is "timeout." Silverstein et al. (1999, p. 7) defined a session by a "series of queries by a single user made within a small range of time." The session is "timed out" after five minutes of user inactivity. The timeout method is commonly used by commercial Web database vendors to gather user statistics, but the length of time varies. Duy and Vaughan (2003) compared vendor data, based on IP addresses and timeouts, with local client-side data, based on login attempts, in a university library. They found that the two data sets had different numbers but similar patterns. Huang, Peng, An, Schuurmans, and Cercone (2003) employed a dynamic session identification method using *n*-gram language models to determine session boundaries. They found that this method performed better than the timeout method. Spink et al. (2001, p. 227) defined a session as the "entire set of queries by the same user over time." The user was identified by an anonymous user code assigned by the Excite server creating the logs. However, it is not clear how the code was assigned, possibly by IP addresses or by cookies, and how the session boundaries were determined. Another study using Excite data (Ross & Wolfram, 2000) revealed that Excite users were identified by cookies. The lack of standards in measurement makes the comparison of results from different studies very difficult. Jansen and Pooch (2001) pointed out this problem and proposed a framework for future studies.

There are other challenges in log file analysis. The log may not record the complete search process when either caching or a proxy server is used. Furthermore, it is difficult to distinguish search queries issued by humans from those of robots. Silverstein et al. (1999) found that robot-initiated queries in the AltaVista logs caused a number of seemingly bizarre results that should be removed prior to analysis. Additionally, it is difficult to determine user demographics such as geographical location. Although this kind of information can sometimes be obtained by an IP address reverse look-up, such data could be misleading because people may register their IP address in other countries for economic or other reasons. To minimize this problem, Nicholas, Huntington, and Williams (2003) proposed a micro-mining method where robust subgroups of users were analyzed. The processing of huge amounts of data contained in a server log file is another challenge. The traditional method of a master file and inverted indexes for text data storage and retrieval may not be sufficiently efficient and flexible. In response, Wang, Berry, and Yang

(2003) have implemented a relational database model that allows them to analyze longitudinal Web server log data with fine granularity. The study identified patterns of querying behavior that provide useful information for Web site design.

Selection and Sampling Methods

The decentralized and dynamic nature of Web pages and Web link creation means that it will not always be possible to find or analyze every page or link; thus sampling may be required. For a tightly defined area, for example, a study of the home pages (only) of the top 100 businesses, it may be possible to retrieve all relevant data. If, however, all business home pages were to be investigated, some kind of random sample would be needed. In any case, where the chosen set of pages would be a subset of the area studied, a potentially difficult problem is the design of an unbiased sample selection method.

Sampling Pages

Computer scientists have designed algorithms to sample pages from the Web randomly. A purely random selection procedure would be one where each page had an equal chance of being selected; but in practice this is not possible or necessarily desirable. A survey of sampling methods is provided by Rusmevichientong, Pennock, Lawrence, and Giles (2001). The general approach is to start with as large a collection of Web pages as possible and then to select from this. Pages are not selected completely at random, but by taking into consideration the link structure of the Web. The underlying model is of a surfer visiting pages partly at random and partly by following links—a “random walk” (Henzinger, Heydon, Mitzenmacher, & Najork, 2000). Algorithms based on this model claim to select popular pages more often than less popular ones, more closely reflecting the actual experiences of Web users. Some of the techniques require the creation of a large collection of pages, from which a random sample can be drawn. Others create the random sample at the same time they build the larger collection. Algorithms that use the second approach can save time and storage space by visiting pages only when necessary for the algorithm; they do not need to download every page in the larger collection. Nevertheless, the set of pages visited must still be much larger than the sample created; researchers wishing to apply this technique to the Web will need access to considerable computing resources.

A commercial search engine could be used in some circumstances. For example, if a sample of pages discussing “Neptune” was needed, a search for the word “Neptune” could be used to produce a list. If the list were too long, the search engine might display only, say, the first 200 matches; but this would be a biased subset, perhaps including those pages judged to be most relevant to the term. Bharat and Broder (1998) have proposed a different search-engine related algorithm for sampling pages. They

use randomly sampled queries created from a Web-based lexicon to generate results pages with which to compare the coverage of different engines. The same basic approach could also be used to generate pages for other purposes. This approach should be less influenced by link structures than random walk methods, although, because search engines tend to use links in their ranking algorithms, there is still likely to be link bias.

Sampling Sites

If there are too many sites to be crawled, some technique must be employed to generate a random sample. If all home page URLs in the total set are known, this problem can be solved with a random number generator (Ju-Pak, 1999). Often, though, random sampling is much more difficult.

Lawrence and Giles (1999a) have used a site sampling method based on randomly sampling Internet Protocol addresses with the aim of generating a random selection of all Web sites. This is no longer recommended since the virtual server capability was introduced, allowing one IP address to host multiple domain names, but only one “main” name. This means that only the main sites run by each Web server could be identified (Fielding, Irvine, Gettys, Mogul, Frystyk, Masinter, et al., 1999; Thelwall, 2002d).

Suppose that the focus of a study is Mexican company Web sites. In order to obtain a representative sample, a complete list would be needed from which to sample with random numbers. Many companies will have domain names ending in .com.mx and administered by a central domain names registry in Mexico. But acquiring a complete list would mean requesting the information from the owning authority. Although direct online access to central domain names databases is a theoretical possibility (Albitz & Liu, 1992), this is not normally permitted. Domain name authorities may be reluctant to give out this kind of information due to fear of spam, although they have done so in the past (Weare & Lin, 2000). One way to circumvent this problem is to randomly generate legal domain names and crawl the associated sites, where they exist (Thelwall, 2002d). For example, a random generator could produce random domain names with all possible combinations between www.aaaa.com.mx and www.zzzz.com.mx (e.g., Thelwall, 2000a). This approach excludes longer domain names (e.g., www.aaaaa.com.mx) and therefore gives a biased sample. If newer Web sites tend to have longer domain names, this would result in age bias. In some circumstances, however, this is the only feasible approach.

This approach presents three other problems: Many Mexican companies probably use other domain names, including .com; some non-Mexican companies probably use .com.mx domains; and some company Web sites will not have their own domain names. All of these problems can be avoided by using an online directory of businesses, such as that in Yahoo!’s Mexican business directories, mx.dir.yahoo.com, but this

introduces a different kind of bias, a human one related to how the sites were identified and selected for the directory.

Probably the most reliable approach would be to sample the companies by name from an up-to-date, official directory and then find their Web sites by asking the companies directly. Alternatively, a search engine search could be used to find the named businesses, although this would introduce search engine coverage and indexing biases.

In summary, there is no fully effective method for sampling business Web sites within a single country. Investigators will have to choose the method that has the source of bias least likely to interfere with the research hypotheses and also record the nature of the bias in their publications.

Large Area Web Studies

Studies that require information about the whole Web currently have three options. The first is to use a crawler to fetch the pages themselves. This is an enormously expensive and time-consuming exercise. The other publicly available option is to access the crawls of more than 10 billion Web pages since 1996 available at the Internet Archive (<http://www.archive.org>), which will require software to process the data. Stanford University's WebBase project (<http://www.diglib.stanford.edu/~testbed/doc2/WebBase>) is similar, but less ambitious in terms of Web coverage. Researchers affiliated with commercial search engines have a third option: to use full search engine crawls. This approach has been used (Broder et al., 2000), but the crawls themselves have remained secret.

Studies that require information about large numbers of Web pages, but not necessarily the full contents of the pages themselves, have an additional option: to use advanced search engine queries to specify the pages to be counted, recording the totals reported. Counts of pages with links from one Web site to another are a common example of the kind of summary statistics that a commercial search engine can yield. However, such results should be treated with caution because of the limited coverage provided by search engines.

Scholarly Communication on the Web

The hope that Web links could be used to provide information similar to that extracted from journal citations has been a key factor in stimulating much webometrics research (Borgman & Furner, 2002; Cronin, 2001; Davenport & Cronin, 2000; Ingwersen, 1998; Larson, 1996; Rodríguez Gairín, 1997; Rousseau, 1997; Thelwall, 2002b). But can this hope be fulfilled? Although structurally very similar to Web links, journal citations are found in refereed documents and, therefore, their production is subject to quality control. They are part of the mainstream of academic endeavor but hyperlinks are not. This caused problems for the early hyperlink-citation analogies, as noted by, for instance, Meyer

(2000); Egghe (2000); van Raan (2001); Björneborn and Ingwersen (2001); and Prime, Bassecoulard, and Zitt (2002). In this section we summarize the results of a series of studies, organized by scale of units analyzed, before considering the underlying motivations for creating links. We conclude with a discussion of how far early hopes have been realized.

The goal underlying almost all of the research reported here is to validate links as a new information source. Such a task entails several different strategies, as reported by Oppenheim (2000) in the related context of validating patent citations. Two of these strategies are to correlate with other data sources of known value and to undertake qualitative studies to uncover creation motivations. With links between university Web sites, for instance, a positive correlation between link counts and research activity would provide some evidence that link creation was not completely random and could be useful for studying scholarly behavior. An important methodological issue is that nonparametric Spearman correlation tests are normally more appropriate than Pearson correlations, given the typically skewed nature of Web link data.

Journals and Scholars

Since much webometrics research has been stimulated by citation analysis, a natural step has been to see if the kinds of bibliometric techniques that are applied to journals and authors can also be applied to e-journals. This could potentially allow a wider range of journals to be studied, giving new insights into scientific publishing. The e-journal is an evolving phenomenon, one that is not yet well understood (Kling & Callahan, 2003).

Smith (1999) pioneered the application of citation analysis to e-journals with a selection of 22 Australasian, refereed e-journals, using AltaVista to count links to them from the rest of the Web. No significant relationship was claimed between these figures and impact factors calculated by the Institute for Scientific Information (ISI). One of the conclusions was that hyperlinks to e-journals are fundamentally different from citations because the former target the whole journal whereas the latter target individual articles. Harter and Ford (2000) reached similar conclusions for a heterogeneous set of 39 journals, also finding no significant correlation between link measures and ISI's impact factors. Vaughan and Hysen (2002) analyzed a discipline-specific set of journals, those from library and information science that were indexed by the ISI. The journals in their study were not e-journals in the sense of providing the full text of articles online, but rather traditional journals with associated Web sites. This time a significant correlation was found between inlink counts and impact factors. The better correlation perhaps reflected greater maturity in the Web. An additional factor may have been that earlier studies were cross-disciplinary, so that disciplinary variations in journal impact factors and Web use obscured underlying trends in the data. Online impact, as measured by total inlink counts, is

therefore associated with offline impact, as measured by average citation counts per article. Why are significant correlations important? They suggest (but do not prove) that hyperlinks may be created for reasons similar to those for citations and therefore may have potential use in bibliometric-style calculations.

Factors not taken into account in the Vaughan and Hysen (2002) study were the content of the journal Web sites and their age. Vaughan and Thelwall (2003) incorporated these two factors for a larger exercise that included 88 law and 38 library and information science (LIS) journals (all ISI journals in the two disciplines that had Web sites). The findings confirmed that in both law and library and information science, counts of links to journal Web sites correlated with impact factors. Unsurprisingly, journals with more online content tended to attract more links, as did older journal Web sites. Evidence was also found that link counts for LIS journals tended to be higher relative to their impact factor than was the case in law. Vaughan and Shaw (2003) took a different approach, comparing citations to journal articles from the ISI's index with citations (not links) to them in the general Web, using Google to collect Web citation data (i.e., how often the journal articles were mentioned in Web pages). All papers published in 1997 in 46 LIS journals were used in this large-scale exercise, which produced predominantly significant correlations, suggesting that online and offline citation impact are in some way similar phenomena. A classification of 854 Web citations indicated that many "represented intellectual impact, coming from other papers posted on the Web (30 percent) or from class readings lists (12 percent)" (Vaughan & Shaw, 2003, p. 1313). Text based citations are more widespread than links to e-journals and are also more directly analogous to bibliometric citations because of the direct connection with individual articles. Data collection is, however, labor intensive. Thelwall (in press a) used this as an opportunity to compare the extent to which pure and applied science journals were invoked in commercial Web sites, but found that in fact businesses rarely mentioned either online. High technology companies were the exception, using published research to support product effectiveness claims.

Citations from the text of online articles rather than links on the general Web have also been objects of study. CiteSeer has been used for this because it maintains an automatically generated database of scholarly publications that are freely available on the Web (Lawrence & Giles, 1999b). CiteSeer incorporates a variety of file formats in addition to HTML. Goodrum, McCain, Lawrence, and Giles (2001) analyzed citation patterns in online PostScript and PDF-formatted computer science papers, finding significant differences when compared with ISI citation data. In particular, conference papers were more frequently cited online (15 percent), but only 3 percent of citations were from ISI computer science papers. This shows the different character of the two data sources. Note that in computer science, conference proceedings are often regarded as equivalent in quality to journals. Zhao and Logan (2002)

conducted a similar study of Extensible Markup Language (XML) research, also using CiteSeer. They recommend that, where possible, online and offline citation analyses should be conducted in parallel, as the data sources did not have the same strengths and weaknesses. Strengths for the Web approach included greater coverage in terms of both number and variety of papers and the inclusion of more information from the full text of the papers. ISI data were more readily accessible, but CiteSeer had the potential for greater automation of data collection techniques, if programming expertise were available to a research team. Most significantly, Lawrence (2001) investigated online computer science research and found that papers publicly available online were more frequently cited (in the ISI database) than those that were not available. It seems likely that placing published articles freely online can generate more interest and facilitate easy access for other scholars. The converse is probably also true; some scholars may choose to publish only their most popular articles online.

The growth of many different types of digital library has created new opportunities for employing quantitative techniques to study aspects of journal article use. We have discussed some research based around the CiteSeer digital library, but there are also public digital libraries that are created by users submitting their articles directly. These are often called preprint archives. Posting research to a preprint archive offers two benefits: It is a public, time-stamped claim to have created research, one that is much faster than journal publication, and it is also a mechanism for dissemination of research. Harnad and Carr (2000) have shown how citations can be turned into links; and in fact the related e-print archive initiative has built in many log file-based tools for analyzing usage patterns. These tools have helped to show that readers in some fields prefer to download complete articles rather than abstracts (Brody, Carr, & Harnad, 2002). Most starkly, more than half of the downloads from the physics archive (<http://arXiv.org>) at any given time are attributable to articles that are less than a week old, and the average elapsed time between depositing a paper and its receiving a citation can be as short as one month.

Web server logs can also be used to give information about the usage of e-journals that are not in digital libraries. Rudner, Gellmann, and Miller-Whitehead (2002) tracked the readers of online education journals using an online survey, server access statistics, and content analysis. Their findings suggested that the readership for this type of journal was much wider and larger than that for print journals. They also found that the most current topics were the most widely read and provided a list of recommended actions for e-journal publishers.

Three studies have chosen scholars rather than journals or journal articles as the basic unit of analysis, collecting data based upon *invocations*: the mentioning of a scholar's name in any context in a Web page. Cronin, Snyder, Rosenbaum, Martinson, and Callahan (1998) found academics to be invoked online in a wide variety of contexts, including

informal documents such as conference information pages and course reading lists. This was used to support a claim that “the Web should help give substance to modes of influence that have historically been backgrounded in the narratives of science” (Cronin et al., 1998, p. 1326) and that it could help to highlight “the diverse ways in which academic influence is exercised and acknowledged” (Cronin et al., 1998, p. 1326). Landes and Posner (2000) used Web invocations as a partial measure of public fame for intellectuals and Cronin and Shaw (2002) pursued this idea for information studies scholars; the population studied had low salience.

National University Systems

Analyzing the interlinking between universities within a single country offers the perfect scale for a study. The number of objects to analyze (one site per university in a country) is manageable and counting all links to a whole university site seems to give a sufficiently high level of aggregation to produce reliable results. In other words, the vagaries of link creation and data collection seem to average out at the level of entire universities, allowing trends to be identified.

Ingwersen (1998) proposed new measures of online impact. The most popular of these, the (External) Web Impact Factor (WIF or Web-IF), was designed to measure the average online impact of a set of Web pages by counting the inlinking pages outside the set in question, and then dividing by the number of pages inside the set. The WIF was later modified for universities by using another measure of university size: its full-time faculty (Thelwall, 2001d). This was justified by reasons that eventually led to the creation of the Alternative Document Models discussed earlier. For example, a university may post huge numbers of pages on its site, such as server log statistics, that are not intended for people outside the university. Server log statistics would be unlikely to attract links from other universities, but may run to tens of thousands of pages and, hence, reduce the average number of links per page. The modified (faculty denominator) WIF counts links per full-time faculty member, avoiding the problem of lower scores from posting large numbers of pages that are unlikely link targets. Most university studies have analyzed interuniversity links, but some have focused instead on connections between universities and other sectors of society, such as industry and government (Leydesdorff & Curran, 2000). If the total impact of university Web sites could be easily measured and compared through hyperlinks, this would clearly be useful information, given the importance of the Web for information seeking.

As was the case with journals, early studies centered on trying to show that link metrics correlated with research measures—a first step toward assessing their similarity to citations (e.g., Smith, 1999; Thelwall, 2000b). Early findings were negative (Smith, 1999; Thelwall, 2000b), and the reasons suggested included failures in the search engines used to

obtain the link counts and the number of links created for reasons unrelated to research. Counts of links to a set of 25 U.K. universities were subsequently found to correlate significantly with average research productivity, (Thelwall, 2001b), although no claim was made that research activity caused link creation. Both AltaVista and a specialist information science Web crawler were used in this study, giving similar results. A comparable relationship was later found for Australia (Smith & Thelwall, 2002) and also Taiwan (Thelwall & Tang, 2003), using different national measures of research productivity. The correlations, although significant, were not very high and graphs of the data showed individual universities that did not fit the pattern. These anomalies occurred when individuals or automated processes had created huge numbers of links. For example, a biochemistry database at Warwick University contained tens of thousands of links to a similar online database at the University of Cambridge, dwarfing other link counts (Thelwall, 2002b).

The discovery of anomalies in link data led to the search for alternative methods that would be more robust, in particular to the development of the ADMs discussed previously. These were applied to the U.K. with an extended data set of 108 institutions, and produced much more significant results, particularly for the directory and domain versions (Thelwall, 2002b). In the experiment, results for the four ADMs were compared with estimated research productivity for the 108 U.K. institutions. Research productivity was calculated from a reliable, peer-review based official source (<http://www.rae.ac.uk>). The underlying hypothesis was that a better ADM would generate results that would correlate more strongly with research productivity figures. Although the Spearman correlation differences were small, the directory and domain graphs were much more linear, giving additional support for a closer relationship between links and research for the directory and domain ADMs. Hybrid ADM counting methods were later developed, where multiple links to a target document from the same university were removed, giving even more significant results (Thelwall & Wilkinson, 2003b). These "range" models used a counting approach based upon each of the four document models, but never counting the same link target more than once for the same source university. For example, if using the domain ADM and domains A and B in one university targeted domain C in another, this would score as only one link in the range model because the same domain was the target. The hypothesis behind the range models was that linking practices are frequently shared within a university, so that multiple links to the same target could cause discrepancies in counts. An example of this is when many pages in one university target a clickable list of U.K. university home pages. There are several of these link lists on the Web, but some universities have many links to just one. Presumably this is the result of imitation within the university.

The ADMs have also been combined with a manual classification scheme. By restricting inlink counts to target pages mainly connected

with research, the Spearman correlations reached a value of 0.949 for the directory range model, compared to 0.940 for the same counting model applied to all target pages and 0.920 for the standard file model applied to all target pages (Thelwall & Harries, 2003). The effect of applying the directory and domain ADMs coincided with restricting the counts to classified pages, indicating that anomalies were predominantly found in pages not targeting academic content. This gives a high degree of confidence that links between university Web sites are connected with scholarly activity in some way, despite the number that are created for recreational reasons. Of course, it still does not establish a cause-and-effect relationship.

The significant correlations found between universities opened the door to attempts to mine deeper for patterns and to model the linking process. Metrics to measure the use of the Web by universities (via outlinks) and the connectivity of a university with its peers were proposed (Thelwall, 2003b). These were termed the Web Use Factor (WUF) and the Web Connectivity Factor (WCF), respectively. It was expected that the WUF would be less reliable than the WIF, being dependent upon the crawling of a single site to identify its outlinks, rather than upon multiple other sites to compile total inlinks. This was not borne out by a detailed examination of the data, however. The WCF was designed to restrict the impact of outliers on the data by assessing the link strength between a pair of universities to be the minimum of the count of links from the first to the second and the count of links from the second to the first. This should remove unidirectional anomalies. However, this more complex measure was not found to be significantly more robust, based upon the data set used (108 U.K. universities). An interesting development was the use of the measurements to provide baselines from which to compare the WIFs, WUFs, and WCFs of individual universities in order to identify those that were not well connected on the Web, perhaps indicating underlying problems in university Web usage or publishing policies.

In a different approach, the linking process has been modeled using link counts between all pairs of 86 U.K. universities as the raw data (excluding universities generating statistical anomalies), and regressing these against combinations of faculty size and average research productivity (Thelwall, 2002e). The university research productivity figures used were estimates derived by multiplying peer-reviewed research quality ratings of universities by total active researchers. This is a hybrid research measure (quality times quantity) but one that has some currency inside the U.K. because of its implicit use in newspaper ranking tables. The peer-review component, the Research Assessment Exercise, is used to direct U.K. government research funding and is very important to all universities. The best predictor of total links between a pair of institutions was the product of their two total research productivities. Combined with the previous study, this gave evidence that outlink creation is not fundamentally different from inlink reception. Both are equivalent in the model.

Four further investigations have given insights into the phenomenon of academic Web site interlinking. First, geographic factors for interlinking were investigated, revealing that the extent of interlinking between pairs of U.K. universities decreased with geographic distance (Thelwall, 2002c). In particular, neighboring institutions were very much more likely to interlink than average. This shows that despite the existence of collaboratories and other devices for virtual collaboration, and their undoubted use for international communication, the Web is not divorced from physical reality. Second, alternative sources of links were investigated, classified by top-level domain. It was found that links to U.K. universities from .edu sources produced results very similar to those between U.K. universities, suggesting that academic link attractiveness is a phenomenon that is numerically similar nationally and internationally (Thelwall, 2002a). Third, an investigation into personal home pages that linked to U.K. universities found that this source of links gave very similar quantitative results to inter-university links, even though almost a third of the links was for recreational purposes (Thelwall & Harries, 2004). It can be concluded that link attractiveness is a relatively robust construct for university Web sites, at least in the U.K., but that it is not exclusively dependent on the academic content of the pages linked to, even though link counts correlate highly with university research productivity. A fourth study compared different potential link attractors for Canadian universities, finding that faculty quality and language were the best predictors of inlink counts (Vaughan & Thelwall, in press a). French language universities in Canada attracted significantly fewer links than comparable English language universities.

The most significant quantitative result was that the volume research produced by academics was the main reason for attracting links: Universities with better researchers attracted more links because the researchers produced more Web content, rather than because the content produced was of a higher “link attractiveness” (Thelwall & Harries, 2004b). This is in contrast to the case for formal scholarly publications, where “better” scholars tend to produce articles that attract more citations (Borgman & Furner, 2002), and is a critical finding suggesting that link counts should not be seen as in any way signifying quality.

International University Comparisons

Studies comparing university Web sites internationally have revealed interesting but fairly predictable descriptive patterns. These studies differ from those described in the previous section by being relational: looking for evidence of international linking trends. Smith and Thelwall (2002) compared linking between British, Australian, and New Zealand universities and found that New Zealand was relatively isolated on the Web, in line with findings from a previous bibliometric study for journals

(Glänzel, 2001). A new measure was introduced here, the propensity to link. This measure was taken to be the total links from all universities in one country to all universities in the second, divided by the total number of faculty in the first country and also by the total number of faculty in the second country. A crawler was used to collect the raw data. A normalizing calculation allowed these figures to be compared with links between universities in the same country. Not surprisingly, international links were dwarfed by national links, even for historically related and traditionally collaborating countries.

A larger follow-up study used AltaVista data and simple network diagrams (Thelwall, 2001a) to map the interlinking between universities in the Asia-Pacific region (Thelwall & Smith, 2002). In the four types of diagrams, the width of arrows between countries was proportional to raw link counts, links divided by target system size (total number of Web pages in all universities in the target country), links divided by source system size, and links divided by source and target system size. All four diagrams were very different, but showed that Australia and Japan were central to the academic link structure of the region, with smaller countries attracting attention disproportionate to their size. The diagrams were most useful when comparing Web collections of approximately the same size.

There have been three published studies of European university Web sites. Polanco, Boudourides, Besagni, and Roche (2001) clustered universities throughout the European Union (E.U.) based upon links. A general survey of E.U. university Web site sizes revealed a huge disparity between the West and East, one that could hinder attempts to use the Internet to integrate European research (Thelwall, Binns, Harries, Page-Kennedy, Price, & Wilkinson, 2002). AltaVista's linguistic capability was employed to examine the languages used for European Web pages and link pages in particular (Thelwall, Tang, & Price, 2003). English was the major language for the whole of the E.U. with the exception of Greece, both for all pages and for internationally linking pages. Beyond Greece and the English speaking nations, English language pages formed approximately half of all international link pages, with the other accounted for by indigenous languages. Not surprisingly, links between countries with a shared language were also common, particularly in the shared language. Apart from English, Swedish was the only other case of a language used extensively for international linking, mainly within Scandinavia; it should be noted that Swedish is very close to other Scandinavian languages.

Departments within a Discipline

In parallel with investigations into interlinking between universities, there have been studies of interlinking between departments within a single discipline. Success with research having a disciplinary focus could potentially lead to the development of techniques to measure subject

communication patterns on the Web, perhaps even giving early indications of emerging interdisciplinary trends. Thomas and Willett (2000) studied U.K. library and information science departments, but found no significant correlation between inlink counts and research ratings. Significant differences between inlink counts and *usnews.com* rankings were found later for U.S. library and information science schools (Chu, He, & Thelwall, 2002). Since then, significant research and inlink correlations have been found for U.K. computer science departments (Li, Thelwall, Musgrove, & Wilkinson, 2003), U.S. psychology and U.S. chemistry departments (Tang & Thelwall, 2003). Interlinking between U.S. history departments was too low for patterns to be extracted, and there were significant disciplinary differences in patterns of interlinking within psychology, chemistry, and history (Tang & Thelwall, 2003). No geographic trends were evident, perhaps either because geography is less important in the U.S. than the U.K. or because the phenomenon is less evident within a single subject. History may actually be an anomaly in the humanities, because other humanities subjects seem to publish more on the Web (Thelwall, Vaughan, Cothey, Li, & Smith, 2003).

There is an important issue of scale when moving from analyzing entire university sites to analyzing individual departments. The studies reported here indicate that interdepartmental link data are sufficiently strong in some disciplines to support an analysis, but the research relationships found have been much weaker than for whole universities. Presumably this is due to the averaging effect of the larger units of analysis—in other words, the much larger number of links between universities than between departments makes trends more apparent. It would be an attractive research question to analyze the interlinking of departments in terms of cluster analysis approaches (analogous to Small's [1999] science mapping), but it is not yet clear that there is enough link data to do this meaningfully. In addition, the initial task of accurately identifying all Web sites associated with a discipline is far from trivial (Li et al., 2003). An alternative approach may be for more text-based analyses to circumvent the problem of link sparseness, despite the greater technical difficulty with this mixed approach.

Scale is also a relevant factor when considering the kinds of institutional analyses common in scientometrics, for example, evaluating all departments within a single university. The relatively small number of inlinks per department, coupled with the inherently skewed nature of Web linking, means that inlink counts can be only very weak indicators of visibility or impact and should certainly be interpreted with great caution. However, the Web is important to research communication, and thus departments attracting a low number of inlinks compared with other departments from the same discipline could reasonably be asked at least to address the issue of whether their Web presence is effective (see Thelwall, 2002f).

Link Creation Motivations

Link creation motivation studies are vital for developing an understanding of how link counts should be interpreted. They have tended to be less popular than statistical studies, although both are needed to validate link counts. That said, most of the correlation-based studies discussed here have included some basic analysis of Web pages in order to investigate factors related to motivations or anomalies (e.g., Thelwall, 2001d, 2002b; Thelwall & Wilkinson, 2003b), some through the employment of typologies to categorize link types. In fact, most motivation studies have investigated the context of links, rather than attempt to directly ascertain author motivations. Motivation studies should be viewed in relation to what is known about Web use in general. It is important to understand that Web use is not determined by technology, it is context specific (Hine, 2000). In particular, academics use the Web in different ways and this is likely to continue (Kling & McKim, 2000).

Kim (2000) investigated authors' motivations for creating links in e-journal articles. Some motivations were similar to those previously discovered for making citations, but others were new. The new motivations related to the accessibility and richness of electronic resources. S. D. Herring (2002) conducted an exploratory citation analysis of e-journal articles and found significant use of electronic resources.

Early studies of links found that many were created for nonacademic reasons, such as to connect to a list of local restaurants (Smith, 1999; Thelwall, 2001d), which undermined the use of link counts to investigate research relationships. Later studies also found links of this type but they were less prevalent (Thelwall, 2001b; Thelwall & Harries, 2003). A study of the 100 academic pages in the U.K. that were the target of most links from other U.K. universities did not find recreational pages, but mainly general-purpose pages, such as university home pages and some departmental pages (Thelwall, 2002g). Two studies separated links that were closely related to research from others, finding that counts based upon the research-related links associated more significantly with research productivity (Thelwall, 2001b; Thelwall & Harries, 2003). The most comprehensive study so far took a random collection of 414 links between U.K. university Web sites (using crawler data), which were then classified by two independent researchers (Wilkinson, Harries, Thelwall, & Price, 2003). The classification was problematic, because of a low level of agreement on categories. Nevertheless, by combining similar categories more reliable ones were formed and it was shown that more than 90 percent of links were related to scholarly activity in some way, including teaching. However, less than 1 percent were judged to be equivalent to journal or conference article citations. Combined with the correlation results, this gives strong evidence that links between university Web sites can be used as evidence of patterns of informal scholarly communication. Of course, this is only one type of informal scholarly communication, and there is no guarantee that other types will follow a similar pattern. One study investigated links to university home pages

(Thelwall, 2003c), finding different reasons for their use even within this subgroup. Some links did not appear to fulfill a navigation function and others served no communication purpose at all. This exercise showed the difficulty in both classifying link motivations and interpreting link counts, a major problem for academic Web link studies.

Conclusions

From the link motivation findings, it seems that link counts (either inlink counts or inlinking page counts) reflect a wide range of informal scholarly communication. Universities that conduct more and better research tend to attract more links, but only because they publish more Web content not because it is of higher quality. University Web link analysis is complex and should not be the basis for a simple online indicator of research or other impact. One practical use of academic link analysis is as a health check on universities (e.g., Thelwall & Aguillo, 2003). For instance, lower than expected WIF scores for individual institutions can be indicators of potential problems with the university Web site or Web publishing policy. A second application is the provision of data to complement other indicators—triangulation. This is probably most suitable for general impact evaluation at present. It is also valid to use link counts between academic spaces as partial measures of online scholarly communication, remembering that link creation reflects a wide variety of informal scholarly communication. Results must be interpreted carefully, however, because there is no reason to believe that hyperlinks would necessarily be representative of Web use, and other kinds of online scholarly communication, such as e-mail, may well have different patterns.

Heimeriks, Hörlesberger, and van den Besselaar (2003) have proposed using links in conjunction with other sources to develop a broader picture of collaboration; this method holds promise for future link research. They incorporated article citations (bibliographic coupling) and project collaborations in order to map communication and collaboration between institutions in specific advanced scientific fields. The methodological framework used also enables the construction of rich pictures with which to explore the social dimensions of science and university-industry-government (triple helix) relationships.

General and Commercial Web Use

Measurements of the Web

Given the Web's dynamic nature, there is a real need for benchmark data on its size and characteristics. These data also have practical implications for information retrieval, management, cataloging, and indexing. Unfortunately, as we shall see, reliable Web size estimates are impossible.

The Web Characterization project conducted by the Office of Research of the Online Computer Library Center, Inc. (OCLC) collects an annual Web sample based upon random IP addresses to analyze trends in the size and content of the Web (OCLC, Office of Research, 2003). The project compiles various statistics such as the number of Web sites, the annual growth rate of the Web, country of origin, language, and the 50 most frequently linked-to sites. The project estimated that the Web contained 9 million sites in 2002. Other estimates are available, but give very different numbers due to differences in sampling methods and operating definitions (e.g., total Web vs. indexable Web). Lawrence and Giles (1999a) used a random sample of IP addresses to estimate the number of servers and the mean number of pages per server. They then estimated the size of the indexable Web based on these two numbers. This approach suffers from the problem that the number of pages per server follows an extremely skewed distribution. The mean is not an appropriate measure, which limits the accuracy of the estimate, a fact acknowledged by the authors. Their definition of the publicly indexable Web was essentially all pages reachable by a robot following links from the home page of a site, excluding pages banned by the Robots Exclusion Protocol (<http://www.robotstxt.org/wc/exclusion.html>) and excluding sites with authentication requirements. However, as mentioned earlier, the advent of the virtual server meant that it was not possible to equate one IP address with one domain name or Web site. Therefore, IP sampling is no longer capable of giving reasonable Web size statistics (Thelwall, 2002d).

Perhaps more relevant and practicable is coverage of the Web by the major search engines. The well-known Search Engine Showdown site (<http://searchengineshowdown.com>) has regular updates on search engine statistics, such as relative size and database overlaps. The statistics on search engine size are based on counting total hits for single word queries (Notess, 2003). Longitudinal studies analyze changes in the Web rather than total size. Koehler (2002), for example, investigated Web page changes and persistence. He followed 361 pages weekly for four years and recorded changes in content and links. He found the half-life of a Web page to be roughly two years and that different types of Web pages (e.g., commercial, educational) behave differently. The limitation of the study is that these pages were randomly sampled from a single search engine, WebCrawler, in December 1996. These pages may not represent Web pages in general.

Web Server Log Analysis

A long tradition in information science involves examining user online search behavior by monitoring or recording searches in order to improve user interfaces or user training. Numerous studies have analyzed how people use traditional information retrieval systems such as Thomson Dialog or library OPACs (Online Public Access Catalogs). Those studies are typically conducted in experimental settings with academic users.

The Web user has much more heterogeneous search interests and habits. The Web server log, which unobtrusively captures real people's natural search processes, is a rich source of insights on Web use. Such information is useful not only for Web design but also for understanding the social impact of the Web.

Despite the difficulties in server log analysis, a significant body of research findings has been built up concerning the use of commercial search engines. Analyses of server logs from two search engines, AltaVista and Excite, reveal similar user search patterns. Web users differ significantly from users discussed in standard IR literature. Most people use short queries, seldom modify the query, rarely use advanced search features, and look at only the first 10 to 20 results (Silverstein et al., 1999; Spink, Jansen, & Ozmutlu, 2000; Spink, Wolfram, Jansen, & Saracevic, 2001). There are large numbers of mistakes or incorrect uses (Jansen, Spink, & Saracevic, 2000). Longitudinal analysis of Excite server logs has shown that search topics have shifted, but there has been little change in user search behaviors (Spink, Jansen, Wolfram, & Saracevic, 2002) even though there was increased use of advanced search features such as Boolean operators and relevance feedback (Wolfram, Spink, Jansen, & Saracevic, 2001). Ross and Wolfram (2000) investigated the topic content of queries submitted to the Excite search engine using a content classification of term co-occurrence within unique queries. This analysis provided a better understanding of the types of topics users were searching for and their relationships. Because unique queries, not specific users, were of interest, the problems of user and session identification were not at issue.

In addition to search engines, studies of the logs of individual Web sites can give useful information about user interactions with electronic sources. This is particularly relevant for information providers, such as health sites. Collating data from their previous studies (e.g., Nicholas, Huntington, Lievesley, & Wasti, 2000; Nicholas, Huntington, & Williams, 2002) and triangulating with questionnaire data, Nicholas, Huntington, Williams, and Dobrowolski (2004) summarized and characterized users' searching behavior as seldom penetrating a site to any depth, tending to visit a number of sites for a given information need, and seldom returning to sites previously visited. The triangulation approach, using different sources of information for the same research goals, is good practice for incorporating user perspectives in the construction of convincing theories of user behavior, especially given the problems of interpreting log files.

Chen and Cooper (2001) used clustering techniques to detect usage patterns without any demographic information in a Web-based library catalog system. Based on the usage groups classified, they further developed a Markov model that can predict users' moves from one search state to the next (Chen & Cooper, 2002). Although these studies were carried out on a Web-based library catalog system, the methods and the models developed could be applied to other Web systems.

A connection between log files and hyperlinks can be made because log files can record the URL of the page previously visited by the user in addition to the URL of the current one. This normally implies that a link on the previous page has been traversed. One team of researchers has pursued a fruitful line of research in which this kind of implied link navigation is used as a measure of the importance of links, or the likelihood of their future use. This approach has allowed them to devise improved tools for the navigation of individual Web sites by clustering pages for navigation interfaces in a way that helps users retrieve information from a site (Zhu, Hong, & Hughes, in press), by predicting paths that users will traverse within a site (Zhu, Hong, & Hughes, 2002), and by adapting PageRank to take into account the frequency of use of links in a site (Zhu, Hong, & Hughes, 2001). Thelwall (2001f) used the same server log information to study site inlinks and found that many generate little or no traffic. This was a novel use of both the inlink data (usually obtained from search engines) and the log file data. The combination of the two data sources provided a fuller picture than either data source alone.

Commercial Web Site Studies

Most webometric studies to date have focused on academic or scholarly Web sites, following in the tradition of bibliometrics. This is ironic given that the Web is dominated by commercial sites. Knowledge of commercial Web sites has become increasingly important with the rapid growth of e-commerce.

Thelwall (2000c) randomly sampled .co.uk domain names to determine the kind of U.K. businesses that were most prominent in the early stages (1999) of business Web use. The study found that media-related businesses and professions (e.g., legal firms) had a strong presence on the Web, although the computing industry remained the dominant publisher. In general, commercial use of the Web at that time was in an experimental phase. This was confirmed by a related international study (Thelwall, 2000a), which found many businesses did not design their Web sites to be search engine friendly. As a result, 23 percent of the sites were not registered at all in the five major search engines tested. About two thirds of the sites did not use HTML meta tags for indexing. Thelwall (2001c) also studied hyperlinks in commercial sites and found that roughly 66 percent did carry links pointing to other related Web sites; most often links were directed to sites with business relationships.

Vaughan and Wu (2003) studied the Web sites of China's top 100 information technology companies and found significant correlations between counts of inlinks to a company's home page and the company's business performance measures, including revenue, profit, and research and development expenses. There is also a statistically significant correlation between the age of a Web site and its inlink counts, with older Web sites having more inlinks. Vaughan (2004) used the same approach to examine the top U.S. information technology companies. Again, the

inlink counts correlate significantly with business measures of revenues and profits. Moreover, the two sets of correlation coefficients for the two countries are remarkably similar, given that the two groups of Web sites have very different characteristics.

Many studies of commercial Web sites aim to discover information for competitive intelligence purposes. There are two approaches for this. One analyzes Web site content and the other focuses on link structures. Examples of the former are Zanasi (1998) and Madnick and Siegel (2002). Liu, Ma, and Yu (2001) proposed and tested a number of automated methods to help users find novel information from competitors' Web sites, for example, unexpected services or products that competitors offer. The authors targeted competitors' Web site content but stated that their future work would use metadata and links to extract additional information. Reid (2003) presented a framework for mining the Web's hyperlink structure to gather business intelligence and demonstrated the method through a case study of MicroStrategy, Inc., a business intelligence vendor. Step one is to identify the implicit business community through a search of links to the MicroStrategy Web site. The Web sites of this community are then content analyzed to detect early warning signals of relevance to MicroStrategy. For example, the Web community may create directories or news stories that link to the MicroStrategy site. If they express their perceptions about the company in these news stories, this may give MicroStrategy signals about its business opportunities and challenges.

Social Network Analysis

Social network analysis, a research approach developed primarily in sociology, social psychology, and communication science, focuses on patterns of relations among people, and among groups such as organizations and states. Computer networks, including the Web, can host social networks because they connect people and organizations. Garton, Haythornthwaite, and Wellman (1999) have argued for the usefulness of social network analysis in the study of computer networks in general and the Internet in particular. They explain some basic concepts of social network analysis, describe data collection methods, and demonstrate their application. Their article is an excellent introduction to social networks analysis and a comprehensive review of the related literature. Garton et al. (1999) discussed the applicability of social network analysis to computer networks in general; Jackson (1997) focused specifically on Web hyperlinks as a communication tool and discussed how to interpret Web link structures through social network analysis. Otte and Rousseau (2002) also pointed out the applicability of social network analysis to information science, specifically to the study of the Internet.

Two successful applications of social network analysis to Web phenomena are by Garrido and Halavais (2003) and Park, Barnett, and Nam (2002), both using links as their data source. The former studied

networks of support for the Zapatista movement in Mexico, in which the Internet played a central role. The authors collected data on links to the Zapatista site and mapped these into a Zapatista network on the Web. This hyperlinked network provided an insight into the character of the Zapatista's success, for example, the strong connection with women's rights. Park et al. (2002) examined South Korea's 152 most popular Web sites. They operationalized affiliation between Web sites with links and applied cluster analysis to reveal the underlying network. Financial Web sites were found to occupy the most central position in the network.

Adamic and Adar (2003) studied home page owners in two U.S. universities to explore how well home page texts and links, as well as e-mail list membership, could predict social connections. The degree of success varied with the precise nature of the data: for example, two students with *co-outlinks* (analogous to bibliographic coupling, cf. Figure 2) to Yahoo! would not be as good an indicator of a likely social connection as *co-inlinks* (co-citation) from another student's "list of friends" would be. The software developed in the study was able to suggest reasons why two individuals could be friends, such as membership of common e-mail lists or shared topics of interest. The main limitation of the study was the atypical nature of the communities studied: students in two elite universities.

Links have been used in conjunction with other data sources to explore online social behavior around specific topics. Hine's (2000) virtual ethnography uses links as one element in a broad analysis of phenomena with online components. Her insistence that online actions should be interpreted in conjunction with offline actions is particularly welcome: The Web is not divorced from the real world. Web Sphere Analysis is another sociological method that has been used to study Web publishing in relation to U.S. Presidential elections (Foot, Schneider, Dougherty, Xenos, & Larsen, 2003). It includes links as part of a wider exploration of online relationships and genres of publishing. Detailed descriptions of both techniques are beyond the scope of this review but both represent thoughtful and appropriate frameworks for studying online phenomena.

Topological Modeling and Mining of the Web

If the Web is viewed as a collection of pages connected by links and all information about page contents, locations and URLs is discarded, then this produces what is called a *directed* graph in mathematics if the direction of the links is kept, or an *undirected* graph if the direction of the links is ignored. Physicists, computer scientists, and information scientists have attempted to build models of the Web or Web growth through link analysis, as well as to extract abstract clustering patterns. These approaches can be termed topological because they treat the Web as a mathematical graph and ignore spatial relationships and page contents. The Web is the largest network for which topological information is currently available, and it has thus become a testing ground for many modeling efforts (Albert & Barabási, 2002). Graph theoretic approaches have

been used in bibliometrics and scientometrics since the 1960s for analyzing citation networks and other information networks (e.g., Furner, Ellis, & Willett, 1996; Garner, 1967; Hummon & Doreian, 1989; Nance, Korfhage, & Bhat, 1972) so it is natural for information scientists to be interested in Web graphs.

Two key algorithms that have served to underline the potential of hyperlink structure analysis are Kleinberg's (1999) HITS (Hypertext-Induced Topic Search) and Google's PageRank (Brin & Page, 1998). A lesser known approach, Constrained Spreading Activation (Cohen & Kjeldsen, 1987), has also had an impact in computer science when applied to links. The purpose of HITS is to identify the most useful Web pages for any given topic. The algorithm identifies relevant pages by text matching but picks out the most relevant through an iterative link-based algorithm that attempts to find "authorities," pages with many inlinks from within the topic, and "hubs," pages with many outlinks to topic pages. Good authorities and hubs are both desirable. The approach has been shown to be adaptable for different purposes (Lempel & Moran, 2002). In a topological analysis of the U.K. academic Web space, Björneborn (2003) found an indication that Web sites that are Kleinbergian hubs and authorities also have high *betweenness centrality* in the investigated Web space. Betweenness centrality is defined as the probability that a node (e.g., a Web site) will occur on a shortest path between two arbitrary nodes in a network graph (Freeman, 1977).

Unlike HITS, PageRank is not topic-centered but ranks all pages using the two assumptions that pages with more inlinks are more likely to be useful and that inlinks from useful pages are more important than inlinks from average pages (Brin & Page, 1998). This iterative algorithm is particularly valuable because its mathematical implementation makes it feasible to implement on billions of Web pages. Logically, PageRank should be a better indicator of value than simple inlink counts, but this is difficult to demonstrate because of the confounding effect of internal site selflinks for navigational purposes (Thelwall, 2003a). Highly ranked pages have hub-like properties (i.e., they point to, rather than directly contain, useful content): in university Web sites they are likely to be the home pages of universities and departments (Thelwall, 2003a).

Network Models of Linking Behavior

Fundamental to an understanding of linking on the Web is the occurrence of many so-called, scale-free network features (Barabási & Albert, 1999; Barabási, Albert, & Jeong, 2000). In a scale-free network there is no "typical" node, that is, no characteristic "scale" to the degree of connectivity (Ball, 2000). Scale-free distributions of inlinks and outlinks show long, power-law tails, meaning that only a small number of Web nodes receives or provides many links, whereas the majority has few inlinks or outlinks. A range of power-law distributions has been identified

on the Web in, for example, scattering of TLDs on a given topic (Rousseau, 1997), inlinks per site (Adamic & Huberman, 2001; Albert, Jeong, & Barabási, 1999), outlinks per site (Adamic & Huberman, 2001), in-neighbors and out-neighbors per site (Björneborn, 2003), pages per site (Adamic & Huberman, 2001; Huberman & Adamic, 1999), visits per site (Adamic & Huberman, 2001; Huberman, Pirolli, Pitkow, & Lukose, 1998; Pitkow, 1998), and visited pages within a site (Huberman, Pirolli, Pitkow, & Lukose, 1998; Pitkow, 1998).

Barabási and Albert (1999) argue that scale-free link distributions are rooted in two generic mechanisms of many real-world networks: *continuous growth* and *preferential attachment* (the “rich-get-richer” phenomenon). In this framework, the Web is an open, self-organizing system that grows by the continuous addition of new nodes and links where the likelihood of connecting to a node depends on the number of links already attached to the node. This significance of preferential attachment for power-law distributions is well known in bibliometrics as the Matthew effect, “unto every that hath shall be given” (Merton, 1968), and “cumulative advantage” (Price, 1976).

Barabási & Albert’s (1999) original model for explaining the emergence of scale-free, power-law properties in networks has been augmented with additional factors, including initial attractiveness (Dorogovtsev, Mendes, & Samukhin, 2000), competition and fitness (Bianconi & Barabási, 2001), optimization (Valverde, Ferrer i Cancho, & Solé, 2002), uniform attachment (Pennock, Flake, Lawrence, Glover, & Giles, 2002), transitive linking (Ebel, Mielsch, & Bornholdt, 2002) and lexical distance (Menczer, 2002). Moreover, Amaral, Scala, Barthélémy, and Stanley (2000) explain deviations in scale-free distributions by factors such as aging, cost, and the capacity constraints of network nodes.

The model presented by Pennock et al. (2002) explains the highly skewed and hooked shape of power law diagrams for linking in terms of competing tendencies for new links to be allocated to already highly linked pages and for links to be allocated at random (in reality, probably to new high quality pages). In fact, the balance of competition varies by domain. When the distribution of links was compared for the same type of page, it was found to be more log-normal than like a power law, with additional variations by type. For example, university home pages attract links following a pattern that is less skewed than scientists’ home pages (i.e., links to university home pages are more equally distributed than links to scientists’ home pages). This implies that scale-free network growth models are not necessarily useful for explaining individual linking behavior.

Overall Web Topology

Various topological properties of the overall structure of the Web have been investigated, which helps us understand how search engines work in practice and how the Web has developed. The first important

contribution was a conjecture that any pair of Web pages would be connected by a short chain of links, with an average of only 19 links (Albert, Jeong, & Barabási, 1999). This was later discovered to be incorrect (Broder et al., 2000) because many pairs of Web pages are not connected at all. However, the small-world phenomenon (Watts & Strogatz, 1998)—where pages tend to cluster in groups but, almost paradoxically, pairs of pages can be connected across cluster boundaries without a tortuously long chain of links—persists in large graph components of the Web (Broder et al., 2000).

The groundbreaking Web topology study was that of Broder et al. (2000), which used two full AltaVista crawls from May and October of 1999 and specially constructed software to process 200 million pages and 1.5 billion links. More than 90 percent of the links formed a huge connected group, provided the direction of the links was ignored. This central component split into four roughly equal parts. The core was a strongly connected component (SCC) in which all pages traced a directed link path to all others in the SCC. The other parts comprised a set of pages (OUT) that could be reached from the SCC by following (directed) links, a set of pages (IN) that could reach the SCC by following (directed) links, and the rest (TENDRILS). The remaining pages that were not connected in any way from the main 90 percent were dubbed DISCONNECTED. A problem with this study is the difficulty in extrapolating from it to the whole Web. AltaVista finds pages partly from user submissions of URLs, but primarily by following links from previously visited pages. As a result, pages that are not well linked to are more likely to be missed by its crawler. Thus DISCONNECTED is likely to be far greater for the whole Web, but it is not possible to estimate its size because there is no practical way to automatically find pages that are not linked to. Nevertheless, the results are interesting because AltaVista presumably covers the most used and most visited sites, and its database presumably reflects the Web that users actually see.

Evidence has been found that similar graph component structures are present in parts of the Web, including national university systems (Björneborn, 2003; Thelwall & Wilkinson, 2003a) and individual countries (Baeza-Yates & Castillo, 2001). Baeza-Yates and Castillo extended Broder's model by finding that some of the components could be subdivided. As a final point, it would be misleading to think of any of the components as homogeneous or made up of coherent sets of pages. For example, OUT will contain many pages without outlinks that are on the same site as SCC pages (Thelwall & Wilkinson, 2003a).

Clustering in the Web

The Web's structure does not allow the identification and extraction of small groups of pages by finding disconnected components, as the preceding discussion reveals; the connected components of the Web graph, such as the SCC can be very large. However, clustering Web pages is

desirable for several reasons. Displaying similar pages close to each other is a useful technique for improving user interfaces (Amento, Hill, Terveen, Hix, & Ju, 1999). Search results can also be improved by fitting pages into the context of their relationships to topics (Kleinberg, 1999) and for direct webometric applications complete sets of pages associated with a given subject are often needed, including for department-based studies, as already discussed. As a result, it is important to investigate how the Web can be decomposed into coherent clusters other than the directed and undirected components. We will cover only topological clustering here, excluding algorithms that include page content analyses, such as Kleinberg's work (1999). In fact Menczer (in press) has shown that pages about the same topic tend to cluster together topologically, and that similar pages tend to interlink.

Traditional approaches to topological clustering consume too much computing time to make them practical for large graphs. A breakthrough was made, however, in the design of a fast algorithm that was able to create a viable topological cluster around any given set of connected pages (Flake, Lawrence, Giles, & Coetzee, 2000, 2002). This design was then adapted and applied to national collections of university Web sites, revealing the existence of rich community structures (Thelwall & Wilkinson, 2003a). An alternative clustering approach has also been applied, using co-inlinks (Polanco et al., 2001), but at the site level. A recent investigation into the feasibility of using links, co-inlinks (co-citations), and co-outlinks (bibliographic couplings) to identify similar pages concluded that the link topology was such that combinations of link structures were not likely to be more successful than links alone at identifying similar sites, but that the combinations would give a greater range of potentially similar sites (Thelwall & Wilkinson, 2004).

It is tempting to propose the use of topological clustering and social network analysis techniques for extracting field or subject interlinking patterns from national academic Webs, but this does not seem to be a promising line of research because of the relative scarcity of links. Essentially the problem is that while two journal articles on very similar topics should have references in common, there is no reason why two researchers, research groups, or departments with similar interests should link to the same pages: Some do, but many do not. Effective field clustering is likely to require text analysis in addition to link analysis.

Small Worlds

An intriguing dimension of Web topologies is the small-world properties of the Web's strongly connected component (SCC) (Broder et al., 2000). It is noteworthy that the originators of the World Wide Web envisaged small-world properties: "Yet a small number of links is usually sufficient for getting from anywhere to anywhere else in a small number of hops" (Berners-Lee & Cailliau, 1990, online). Small-world Web topologies are concerned with core information science issues such

as the navigability and accessibility of information across vast document networks. For instance, short distances along link paths affect the speed and exhaustivity with which Web crawlers can reach and retrieve pages.

Small-world theory stems from research in Social Network Analysis on short distances between two arbitrary persons through intermediate chains of acquaintances. In a famous study, Milgram (1967) asked a sample of individuals in the U.S. midwest to forward a letter to a person not known to them on the east coast. The letter was to be mailed to a personal acquaintance (on a first-name basis) likely to know the target person. For the completed chains, a median length of five intermediate acquaintances was needed to forward the letter to the target person. This result contributed to the popularized notion of “six degrees of separation” between any two persons in the world through intermediate chains of acquaintances. Manfred Kochen, a visionary information scientist, was a pioneer in small-world theorizing (Kochen, 1989; Pool & Kochen, 1978). Another innovative information scientist and bibliometric pioneer, Eugene Garfield (1979), envisioned that small-world approaches could be used to identify gatekeepers and “invisible colleges” in informal scholarly communication networks. On the Web, small-world topologies could reflect such things as cross-disciplinary contacts and research fronts in the evolving interconnectedness of science (Björneborn, 2001, 2003; Björneborn & Ingwersen, 2001). In detailed case studies of the Web links, pages, and sites that function as connectors across dissimilar topical domains in the U.K. academic Web space, Björneborn (2003) found that personal link creators such as researchers and students were frequent providers of transversal (cross-topic) links, thus creating small-world phenomena in this Web space. These indicative findings also suggest that people and Web sites related to computer science are important cross-topic connectors. The role of computer science for short link distances in an academic Web space most likely reflects the auxiliary function of computer science in many different scientific disciplines, combined with a more experienced and unconstrained Web presence by many persons and institutions in computer science (Björneborn, 2003).

In a seminal paper, Watts and Strogatz (1998) introduced a small-world network model characterized by highly clustered nodes (high average clustering coefficient) as in regular graphs, yet with short characteristic path lengths between pairs of nodes as in random graphs. Watts and Strogatz showed that in a small-world network it is sufficient for a very small percentage of long-range links to function as short cuts connecting distant nodes of the network. This revival of small-world theory triggered an avalanche of research in a wide range of scientific domains including the Web (Adamic, 1999; Albert, Jeong, & Barabási, 1999; Barabási, 2001). The emergence of small-world topologies on the Web and in other complex networks can be attributed to the *scale-free* network features previously discussed. According to Albert & Barabási (2002), a heterogeneous scale-free topology is very efficient in bringing

network nodes close to each other. Small-world structures may thus arise from a scale-free organization in which a relatively small number of well-connected nodes serve as hubs (Steyvers & Tenenbaum, 2001). Well-connected Web nodes often have multiple memberships providing strong and weak ties (cf. Granovetter, 1973) across many topical clusters. Such far-reaching bridging ties function as transversal links, shrinking the distances in a network to form a small world.

To help paint a mental picture of a small-world “crumpled-up” Web space, every new outlink on the Web may be regarded as a hook able to reshape the existing Web network (Björneborn, 2003). By pulling distant Web nodes close to each other (and thereby drawing together the neighborhoods and clusters to which these nodes already belong), transversal links may “crumple up” Web topologies and thus reinforce the small-world nature of the Web.

The scale-free properties of the Web imply a fractal structure where cohesive subregions display the same characteristics as the Web at large (Dill, Kumar, McCurley, Rajagopalan, Sivakumar, & Tomkins, 2001). Such scale-free properties are often regarded as a fingerprint of self-organization (van Raan, 2000). Analyses of the Web show a remarkable degree of self-organization in the shape of clustered hyperlink structures that reflect topic-focused interest communities (Flake et al., 2002; Gibson, Kleinberg, & Raghavan, 1998; Kleinberg & Lawrence, 2001; Kumar, Raghavan, Rajagopalan, & Tomkins, 1999). The coincidence of high local clustering and short global separation (Watts, 1999) means that small-world networks simultaneously consist of small local and global distances, leading to high efficiency in propagating information on both a local and global scale (Marchiori & Latora, 2000). However, Web links do not *directly* channel information flows as social networks, neural networks, or computer networks do. Web links *indirectly* reflect information diffusion among link creators, because added or removed links may reflect changes in topical interests and the social preferences of link creators. On an aggregated, macro level, such dynamic link adaptations could reflect cultural and social currents and formations (Björneborn, 2003).

From a microlevel perspective, the emergence of small-world properties on the Web is associated with decentralized Web page and link creations resulting in topic drift (Bharat & Henzinger, 1998) and genre drift (Björneborn, 2003). Diverse topics and genres of Web pages may be separated by one or just a few shortcut links. Web genres and topics thus exist side by side on the Web. As noted earlier, most links connect Web pages containing cognate topics (Davison, 2000). This topical propensity (Björneborn, 2003) leads to the emergence of topic-focused Web clusters. Such topical clusters may consist of diverse page genres; for instance, in an academic Web space it could be institutional home pages, link lists, course pages, conference pages, personal home pages, link lists, and others, interconnected by links in a *web of genres* (Björneborn, 2003). Hence, topical clusters with genre diversity entail genre drift along

intra-cluster links. At the same time, some Web page genres, for example institutional or personal link lists, are more diversity-prone and thus function as providers of topic drift along *inter*-cluster links. According to Björneborn (2003), small-world phenomena on the Web may thus be created by complementarities of topic drift and genre drift.

The self-organized small-world architecture of large regions on the Web may be an important, nonengineered, organizing principle for structuring a vast information space (Björneborn, 2003). Webometric approaches may reveal characteristics of how such complex, bottom-up-aggregated Web structures provide traversal options and access points to information. In this context, scholarly Web activities are of special interest when one considers science as a complex, largely self-organizing, sociocognitive system (e.g., Leydesdorff, 2001; Sandstrom, 2001; van Raan, 2000). The Web with its capability to host self-organizing activities in the shape of “collaborative weaving” may thus be seen as a natural environment for scholars (Björneborn, 2003, p. 205).

Summary and Concluding Remarks

We have reviewed many different quantitative studies of the Web, concentrating on those with an information science background. A number of general conclusions can now be reached. First, large amounts of data are required for reliable studies, so investigations normally utilize special software or programs such as Web log file analyzers or commercial search engines. This is in the tradition of bibliometrics, with its use of ISI software or specifically designed programs to analyze large databases. The need to process large data sets is not new, but perhaps represents a solidifying trend in information science research.

Another theme is that the quantitative techniques require complementary qualitative methods (user studies, link creation motivation studies) to understand the results completely. A future trend may be to collaborate with other fields such as cultural studies, computer-mediated communication, social network analysis, and social informatics in order to gain the benefit of more qualitative techniques. The motivation and small world findings indicate that interpretations are likely to be complex and difficult to categorize, either on a small or large scale.

Given the many correlation studies reported here, it is important to bear in mind that significant correlations between link counts and research measures are not ends in themselves. The (partial) information they provide relates to the reliability and validity of using link data, but does not tell us how link structure information can be used for other purposes such as assessing the online impact of Web sites. Counts of links to academic Web sites can now be interpreted with some confidence as reflecting a range of informal scholarly communications. Inlink counts primarily indicate the productivity of the target institution, rather than the average quality or impact of the information presented. The most useful information to be gained from this relates to

outliers: the identification of universities that appear to be making particularly good or poor use of the Web. One logical next step from correlations is to look for patterns in the data, perhaps clustering similar topics across different sites or employing social network analysis techniques. It may be that links alone will not be sufficient, however, since interuniversity links are too sparse to expect similar pages or even similar domains to be frequently interlinked. The future may therefore lie in combining link analysis with text analysis; it is likely that considerable additional computing power and expertise will be required for this development. As a result, a promising new direction for webometrics is to combine with computer science to build tools to harness our knowledge of Web phenomena and produce useful information: in other words, Web intelligence and Web mining. Applied to academic Web sites, techniques from these two fields, together with webometrics, form the emerging field of Scientific Web Intelligence (Thelwall, in press c).

Topological approaches to the Web are useful to give a big picture with which to set other link analyses in context. The future for information science research in this area seems to lie in building a more concrete understanding of linking phenomena by tracking individual links, as exemplified by the small-world approach. Links could then be utilized in social informatics for exploring networked knowledge creation and diffusion, including emerging cross-cultural formations; in Web mining for identifying fertile scientific areas for cross-disciplinary exploration; by search engine designers for more exhaustive Web traversal and harvesting, as well as improved visualization and navigation facilities for search results, perhaps stimulating serendipitous browsing.

The Web is continuously changing and developing, which will provide further research challenges and opportunities. The Semantic Web (Berners-Lee & Hendler, 2001) may eventually make the content of Web pages and the connections among them more explicit and transparent to automatic processes. This should open the door for more powerful and robust webometric analysis. It remains to be seen how far the Semantic Web will penetrate, however, and given the past history of the Web it will vary across user communities and its adoption will depend upon the ways in which the Web is able to provide valuable services for specific needs.

Other Web changes are social in nature. As the Web spreads to wider sections of the population of the richer countries and its use becomes more common in poorer nations, new uses can be expected. The Web is, therefore, likely to increase in importance for most aspects of human activity. A challenge and important future direction for webometrics will be to produce tools and techniques that are able to identify, monitor, and partially quantify these developments. These tools will be most useful if they can be made widely available to others studying the Web, including social scientists.

References

- Abraham, R. H. (1996). *Webometry: Measuring the complexity of the World Wide Web*. Santa Cruz, CA: Visual Math Institute, University of California at Santa Cruz. Retrieved November 20, 2003, from <http://www.ralph-abraham.org/vita/redwood/vienna.html>
- Adamic, L., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, 25(3), 211–230.
- Adamic, L. A., & Huberman, B. A. (2001). The Web's hidden order. *Communications of the ACM*, 44(9), 55–59.
- Aguillo, I. F. (1998). STM information on the Web and the development of new Internet R&D databases and indicators. *Online Information 98: Proceedings*, 239–243.
- Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47–97.
- Albert, R., Jeong, H., & Barabási, A. L. (1999). Diameter of the World-Wide Web. *Nature*, 401, 130–131.
- Albitz, P., & Liu, C. (1992). *DNS and BIND*. Sebastopol, CA: O'Reilly.
- Almind, T. C., & Ingwersen, P. (1996). *Informetric analysis on the World Wide Web: A methodological approach to "internetometrics"* (CIS Report 2). Copenhagen, Denmark: Centre for Informetric Studies, Royal School of Library and Information Science.
- Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to "webometrics." *Journal of Documentation*, 53(4), 404–426.
- Amaral, L. A. N., Scala, A., Barthélémy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21), 11149–11152.
- Amento, B., Hill, W., Terveen, L., Hix, D., & Ju, P. (1999). An empirical evaluation of user interfaces for topic management of Web sites. *Conference on Human Factors in Computing Systems*, 552–559.
- Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology*, 1(1), 2–43.
- Baeza-Yates, R., & Castillo, C. (2001). Relating Web characteristics with link based Web page raking. *Proceedings of SPIRE 2001*, 21–32.
- Ball, P. (2000, Oct 25). The art of networking. *Nature Science Update*. Retrieved November 20, 2003, from <http://www.nature.com/nsu/001026/001026-7.html>
- Bar-Ilan, J. (1997). The "mad cow disease," Usenet newsgroups and bibliometric laws. *Scientometrics*, 39(1), 29–55.
- Bar-Ilan, J. (1999). Search engine results over time: A case study on search engine stability. *Cybermetrics*, 2/3(1). Retrieved November 20, 2003, from <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>
- Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes: A review and analysis. *Scientometrics*, 50(1), 7–32.
- Bar-Ilan, J. (2004). The use of Web search engines in information science research. *Annual Review of Information Science and Technology*, 38, 231–288.
- Bar-Ilan, J., & Peritz, B. (2002). Informetric theories and methods for exploring the Internet: An analytical survey of recent research literature. *Library Trends*, 50(3), 371–392.
- Barabási, A. L. (2001, July). The physics of the Web. *Physics World*. Retrieved November 20, 2003, from <http://physicsweb.org/article/world/14/7/9>
- Barabási, A. L. (2002). *Linked: The new science of networks*. Cambridge, MA: Perseus.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512.
- Barabási, A. L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: The topology of the World-Wide Web. *Physica A*, 281, 69–77.

- Bergman, M. K. (2000). *The deep Web: Surfacing hidden value*. Retrieved November 20, 2003, from <http://128.121.227.57/download/deepwebwhitepaper.pdf>
- Berners-Lee, T. (1999). *Web architecture from 50,000 feet*. Retrieved November 20, 2003, from <http://www.w3.org/DesignIssues/Architecture.html>
- Berners-Lee, T., & Cailliau, R. (1990). *WorldWideWeb: Proposal for a hypertext project*. Retrieved November 20, 2003, from <http://www.w3.org/Proposal.html>
- Berners-Lee, T., & Hendler, J. (2001). Publishing on the semantic Web. *Nature*, 410, 1023–1024.
- Bharat, K., & Broder, M. R. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems*, 30, 379–388.
- Bharat, K., & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyper-linked environment. *Proceedings of the ACM SIGIR 21st Annual International Conference on Research and Development in Information Retrieval*, 104–111.
- Bianconi, G., & Barabási, A. L. (2001). Competition and multiscaling in evolving networks. *Europhysics Letters*, 54, 436–442.
- Björneborn, L. (2001). Small-world linkage and co-linkage. *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, 133–134.
- Björneborn, L. (2003). *Small-world link structures across an academic Web space: A library and information science approach*. Unpublished doctoral dissertation, Royal School of Library and Information Science, Copenhagen, Denmark.
- Björneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1), 65–82.
- Björneborn, L., & Ingwersen, P. (in press). Towards a basic framework of webometrics. *Journal of the American Society for Information Science and Technology*.
- Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology* 36, 3–72.
- Bossy, M. J. (1995). *The last of the litter: "Netometrics."* Retrieved November 20, 2003, from <http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2bossy.html>
- Boutell, T. (2003). *What is the World Wide Web?* Retrieved November 20, 2003, from <http://www.boutell.com/newfaq/basic/web.html>
- Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. *Computer Networks*, 33(1–6), 309–320.
- Brody, T., Carr, L., & Harnad, S. (2002). Evidence of hypertext in the scholarly archive. *Proceedings of ACM Hypertext 2002*, 74–75.
- Brookes, B. C. (1990). Biblio-, sciento-, infor-metrics??? What are we talking about? In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90: Selection of papers submitted for the Second International Conference on Bibliometrics, Scientometrics and Informetrics* (pp. 31–43). Amsterdam: Elsevier.
- Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., et al. (1999). Mining the Web's link structure. *IEEE Computer*, 32(8), 60–67.
- Chakrabarti, S., Joshi, M. M., Punera, K., & Pennock, D. M. (2002). The structure of broad topics on the Web. *Proceedings of the WWW2002 Conference*. Retrieved November 20, 2003, from <http://www2002.org/CDROM/refereed/338>
- Chen, C., Newman, J., Newman, R., & Rada, R. (1998). How did university departments interweave the Web: A study of connectivity and underlying factors. *Interacting with Computers*, 10(4), 353–373.
- Chen, H., & Chau, M. (2004). Web mining: Machine learning for Web applications. *Annual Review of Information Science and Technology*, 38, 289–329.

- Chen, H.-M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11), 888–904.
- Chen, H.-M., & Cooper, M. D. (2002). Stochastic modeling of usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*, 53(7), 536–548.
- Chi, E. H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., & Card, S. K. (1998). Visualizing the evolution of Web ecologies. *ACM SIGCHI Conference on Human Factors in Computing Systems*, 400–407.
- Chu, H., He, S., & Thelwall, M. (2002). Library and information science schools in Canada and USA: A webometric perspective. *Journal of Education for Library and Information Science* 43(2), 110–125.
- Cohen, P. R., & Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, 23(2), 255–268.
- Cothey, V. (in press). Web-crawling reliability. *Journal of the American Society for Information Science and Technology*.
- Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on Web-based citation analysis. *Journal of Information Science*, 27(1), 1–7.
- Cronin, B., & Shaw, D. (2002). Banking (on) different forms of symbolic capital. *Journal of the American Society for Information Science and Technology*, 53(14), 1267–1270.
- Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science*, 49(14), 1319–1328.
- Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the World Wide Web. *Information Society*, 16(3), 201–215.
- Davenport, E., & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A Festschrift in honor of Eugene Garfield* (pp. 517–534). Medford, NJ: Information Today Inc.
- Davison, B. D. (2000). Topical locality in the Web. *Proceedings of the ACM SIGIR 23rd Annual International Conference on Research and Development in Information Retrieval*, 272–279.
- Dhyani, D., Keong W., & Bhowmick, S. S. (2002). A survey of Web metrics. *ACM Computing Surveys*, 34(4), 469–503.
- Dill, S., Kumar, S. R., McCurley, K., Rajagopalan, S., Sivakumar, D., & Tomkins, A. (2001). Self-similarity in the Web. *Proceedings of the 27th International Conference on Very Large Data Bases*, 69–78.
- Dodge, M. (1999). *The geography of Cyberspace*. (CASA Working Paper 8). London: Centre for Advanced Spatial Analysis, University College London. Retrieved November 20, 2003, from <http://www.casa.ucl.ac.uk/cyberspace.pdf>
- Dodge, M., & Kitchin, R. (2001). *Mapping cyberspace*. London: Routledge.
- Dorogovtsev, S. N., Mendes, J. F. F., & Samukhin, A. N. (2000). Structure of growing networks with preferential linking. *Physical Review Letters*, 85, 4633–4636.
- Duy, J., & Vaughan, L. (2003). Usage data for electronic resources: A comparison between locally-collected and vendor-provided statistics. *Journal of Academic Librarianship*, 29(1), 16–22.
- Ebel, H., Mielsch, L.-I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E*, 66, 035103(R).
- Egghe, L. (2000). New informetric aspects of the Internet: Some reflections—many problems. *Journal of Information Science*, 26(5), 329–335.
- Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

- Ellis, D., Oldridge, R., & Vasconcelos, A. (2004). Electronic communication and electronic communities. *Annual Review of Information Science and Technology*, 38, 145–186.
- Etzioni, O. (1996). The World-Wide Web: Quagmire or gold mine? *Communications of the ACM*, 39(11), 65–68.
- Fielding, R., Irvine, U. C., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., et al. (1999). *Hypertext Transfer Protocol -- HTTP/1.1*. Retrieved December 12, 1999, from <ftp://ftp.isi.edu/in-notes/rfc2616.txt>
- Finholt, T. (2002). Collaboratories. *Annual Review of Information Science and Technology*, 36, 73–107.
- Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2000). Efficient identification of Web communities. *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining*, 150–160.
- Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of Web communities. *IEEE Computer*, 35, 66–71.
- Foot, K., Schneider, S., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral Web sphere. *Journal of Computer Mediated Communication*, 8(4). Retrieved December 9, 2003, from <http://www.ascusc.org/jcmc/vol8/issue4/foot.html>
- Freeman, L. C. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, 40(1), 35–41.
- Furner, J., Ellis, D., & Willett, P. (1996). The representation and comparison of hypertext structures using graphs. In M. Agosti & A. F. Smeaton (Eds.), *Information retrieval and hypertext* (pp. 75–96). Boston: Kluwer.
- Garfield, E. (1979). It's a small world after all. *Current contents*, 43, 5–10.
- Garner, R. (1967). A computer oriented, graph theoretic analysis of citation index structures. In B. Flood (Ed.), *Three Drexel information science research studies* (pp. 3–46). Philadelphia: Drexel Press. Retrieved November 20, 2003, from <http://www.garfield.library.upenn.edu/rgarner.pdf>
- Garrido, M., & Halavais, A. (2003). Mapping networks of support for the Zapatista movement: Applying Social Network Analysis to study contemporary social movements. In M. McCaughey & M. Ayers (Eds.), *Cyberactivism: Online activism in theory and practice* (pp. 165–184). New York: Routledge.
- Garton, L., Haythornthwaite, C., & Wellman, B. (1999). Studying online social networks. In S. Jones (Ed.), *Doing Internet research: Critical issues and methods for examining the Net* (pp. 75–105). Thousand Oaks, CA: Sage Publications.
- Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 225–234.
- Girardin, L. (1996). Mapping the virtual geography of the World-Wide Web. *Proceedings of the 5th WWW Conference*. Retrieved November 20, 2003, from <http://www.girardin.org/luc/cgv/www5/index.html>
- Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69–115.
- Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing & Management*, 37(5), 661–676.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380.
- Harnad, S., & Carr, L. (2000). Integrating, navigating, and analysing open eprint archives through open citation linking (the OpCit project). *Current Science*, 79(5), 629–638.

- Harter, S. P., & Ford, C. E. (2000). Web-based analysis of E-journal impact: Approaches, problems, and issues. *Journal of the American Society for Information Science*, 51(13), 1159–1176.
- Heimeriks, G., Hörlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2), 391–413.
- Henzinger, M. (2001). Hyperlink analysis for the Web. *IEEE Internet Computing*, 5(1), 45–50.
- Henzinger, M. R., Heydon, A., Mitzenmacher, M., & Najork, M. (2000). On near-uniform URL sampling. *Proceedings of the 9th International World Wide Web Conference, Computer Networks*, 33(1–6), 295–308. Retrieved November 20, 2003, from <http://www9.org/w9cdrom/88/88.html>
- Hernández-Borges, A. A., Pareras, L. G., & Jiménez, A. (1997). Comparative analysis of pediatric mailing lists on the Internet. *Pediatrics*, 100(2), e8. Retrieved November 20, 2003, from <http://www.pediatrics.org/cgi/content/full/100/2/e8>
- Herring, S. C. (2002). Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology*, 36, 109–168.
- Herring, S. D. (2002). Use of electronic resources in scholarly electronic journals: A citation analysis. *College & Research Libraries*, 63(4), 334–340.
- Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible Web crawler. *World Wide Web*, 2, 219–229.
- Hine, C. (2000). *Virtual ethnography*. London: Sage.
- Huang, X., Peng, F., An, A., Schuurmans, D., & Cercone, N. (2003). Session boundary detection for association rule learning using n-gram language models. *Proceedings of 16th Canadian Conference on Artificial Intelligence*, 237–251.
- Huberman, B. A. (2001). *The laws of the Web: Patterns in the ecology of information*. Cambridge, MA: MIT Press.
- Huberman, B. A., & Adamic, L. A. (1999). Growth dynamics of the World-Wide Web. *Nature*, 401, 131.
- Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E., & Lukose, R. M. (1998). Strong regularities in World Wide Web surfing. *Science*, 280, 95–97.
- Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11, 39–63. Retrieved November 20, 2003, from <http://www.garfield.library.upenn.edu/papers/hummondoreian1989.pdf>
- Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236–243.
- Jackson, M. H. (1997). Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication*, 3(1). Retrieved November 20, 2003, from <http://www.ascusc.org/jcmc/vol3/issue1/jackson.html>
- Jansen, B. J., & Pooch, U. (2001). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*, 52(3), 235–246.
- Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing & Management*, 86(2), 207–227.
- Ju-Pak, K. H. (1999). Content dimensions of Web advertising: A cross-national comparison. *International Journal of Advertising*, 18(2), 207–231.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25.
- Kim, H. J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science*, 51(10), 887–899.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604–632.

- Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). The Web as a graph: Measurements, models, and methods. *Lecture Notes on Computer Science*, 1627, 1–18.
- Kleinberg, J., & Lawrence, S. (2001). The structure of the Web. *Science*, 294, 1849–1850.
- Kling, R., & Callahan, E. (2003). Electronic journals, the Internet, and scholarly communication. *Annual Review of Information Science and Technology*, 37, 127–177.
- Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences in the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science*, 51(14), 1306–1320.
- Kochen, M. (Ed.). (1989). *The small world*. Norwood, NJ: Ablex.
- Koehler, W. (2002). Web page change and persistence: A four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2), 162–171.
- Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations*, 2(1), 1–15.
- Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Proceedings of the 8th International World Wide Web Conference*, 403–415.
- Landes, W. M., & Posner, R. A. (2000). Citations, age, fame, and the Web. *Journal of Legal Studies*, 29(1), 319–344.
- Larson, R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of Cyberspace. *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, 71–78. Retrieved November 20, 2003, from <http://sherlock.berkeley.edu/asis96/asis96.html>
- Lavoie, B., & Nielsen, H. F. (1999). *Web characterization terminology & definitions sheet*. Retrieved November 20, 2003, from <http://www.w3.org/1999/05/WCA-terms>
- Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*, 411(6837), 521.
- Lawrence, S., & Giles, C. L. (1999a). Accessibility and distribution of information on the Web. *Nature*, 400, 107–110.
- Lawrence, S., & Giles, C. L. (1999b). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6), 67–71.
- Lempel, R., & Moran, S. (2002). Introducing regulated bias into co-citation ranking schemes on the Web. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, 425–435.
- Leydesdorff, L. (2001). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communication* (2nd ed). Parkland, FL: Universal Publishers.
- Leydesdorff, L., & Curran, M. (2000). Mapping university-industry-government relations on the Internet: The construction of indicators for a knowledge-based economy. *Cybermetrics*, 4(1). Retrieved November 20, 2003, from <http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html>
- Li, X. (2003). A review of the development and application of the Web Impact Factor. *Online Information Review*, 27(6), 407–417.
- Li, X., Thelwall, M., Musgrove, P., & Wilkinson, D. (2003). The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE (Research Assessment Exercise) ranking in 2001. *Scientometrics*, 57(2), 239–255.
- Liu, B., Ma, Y., & Yu, P. S. (2001). Discovering unexpected information from your competitors' Web sites. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Retrieved December 9, 2003, from <http://www.comp.nus.edu.sg/~liub/publications/kdd2001WebComp.ps>
- Madnick, S., & Siegel, M. (2002). Seizing the opportunity: Exploiting Web aggregation. *MIS Quarterly Executive*, 1(1), 35–46.

- Marchiori, M., & Latora, V. (2000). Harmony in the small-world. *Physica A*, 285(3-4), 539-546.
- Matzat, U. (1998). Informal academic communication and scientific usage of Internet discussion groups. *Proceedings IRISS '98 International Conference*. Retrieved November 20, 2003, from <http://sosig.ac.uk/iriss/papers/paper19.htm>
- McKiernan, G. (1996). *CitedSites(sm): Citation indexing of Web resources*. Retrieved November 20, 2003, from <http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm>
- Menczer, F. (2002). Growing and navigating the small world Web by local content. *Proceedings of the National Academy of Sciences*, 99(22), 14014-14019.
- Menczer, F. (in press). Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology*.
- Merton, R. K. (1968). The Matthew effect in science. *Science*, 159(3810, January 5), 56-63.
- Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines: Fluctuations in document accessibility. *Journal of Documentation*, 57(5), 623-651.
- Meyer, E. K. (2000). Web metrics: Too much data, too little analysis. In D. Nicholas & I. Rowlands (Eds.), *The Internet: Its impact and evaluation: Proceedings of an international forum held at Cumberland Lodge, Windsor Park, 16-18 July 1999* (pp. 131-144). London: Aslib/IMI.
- Milgram, S. (1967). The small-world problem. *Psychology Today*, 1(1), 60-67.
- Molyneux, R. E., & Williams, R. V. (1999). Measuring the Internet. *Annual Review of Information Science and Technology*, 34, 287-339.
- Nance, R. E., Korfhage, R. R., & Bhat, U. N. (1972). Information networks: Definitions and message transfer models. *Journal of the American Society for Information Science*, 23(4), 237-247.
- Nicholas, D., Huntington, P., Lievesley, N., & Wasti, A. (2000). Evaluating consumer Website logs: A case study of the Times/The Sunday Times Website. *Journal of Information Science*, 26(6), 399-411.
- Nicholas, D., Huntington, P., & Williams, P. (2002) Evaluating metrics for comparing the use of Web sites: Case study of two consumer health Web sites. *Journal of Information Science*, 28(1), 63-75.
- Nicholas, D., Huntington, P., & Williams, P. (2003). Micro-mining log files: A method for enriching the data yield from Internet log files. *Journal of Information Science*, 29(5), 401-414.
- Nicholas, D., Huntington, P., Williams, P. & Dobrowolski, T. (2004). Re-appraising information seeking behaviour in a digital environment: Bouncers, checkers, returnees and the like. *Journal of Documentation*, 60(1), 24-43.
- Notess, G. (2003). *Search engine statistics: Relative size showdown*. Retrieved November 20, 2003, from <http://searchengineshowdown.com/stats/size.shtml>
- OCLC. Office of Research. (2003). *Web characterization*. Retrieved November 20, 2003, from <http://wcp.oclc.org>
- Oppenheim, C. (2000). Do patent citations count? In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A Festschrift in honor of Eugene Garfield* (pp. 405-432). Medford, NJ: Information Today.
- Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6), 441-454.
- Park, H. W., Barnett, G. A., & Nam, I. (2002). Hyperlink-affiliation network structure of top Web sites: Examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science and Technology*, 53(7), 592-601.
- Park, H. W., & Thelwall, M. (2003). Hyperlink analysis: Between networks and indicators. *Journal of Computer-Mediated Communication*, 8(4). Retrieved December 9, 2003, from <http://www.ascusc.org/jcmc/vol8/issue4/park.html>

- Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Sciences*, 99(8), 5207–5211.
- Peres Vanti, N. A. (2002). Da bibliometria à webometria: Uma exploração conceitual dos mecanismos utilizados para medir a informação e o conhecimento. *Ciência Da Informação*, 31(2), 152–162.
- Pirolli, P., Pitkow, J., & Rao, R. (1996). Silk from a sow's ear: Extracting usable structures from the Web. *ACM SIGCHI Conference on Human Factors in Computing Systems*. Retrieved November 20, 2003, from http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html
- Pitkow, J. (1998). Summary of WWW characterizations. *Computer Networks and ISDN Systems*, 30(1–7), 551–558.
- Polanco, X., Boudourides, M. A., Besagni, D., & Roche, I. (2001). *Clustering and mapping Web sites for displaying implicit associations and visualising networks*. Patras, Greece: University of Patras. Retrieved November 20, 2003, from http://www.math.upatras.gr/~mboudour/articles/Web_clustering&mapping.pdf
- Pool, I. d. S., & Kochen, M. (1978). Contacts and influence. *Social Networks*, 1, 5–51.
- Price, D. J. D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5), 292–306.
- Prime, C., Bassecoulard, E., & Zitt, M. (2002). Co-citations and co-sitations: A cautionary view on an analogy. *Scientometrics*, 54(2), 291–308.
- Rasmussen, E. (2003). Indexing and retrieval for the Web. *Annual Review of Information Science and Technology*, 37, 91–124.
- Rehm, G. (2002). Towards automatic Web genre identification. *Proceedings of the 35th Hawaii International Conference on System Sciences*. Retrieved January 9, 2003, from <http://www.computer.org> (electronic restricted access document).
- Reid, E. (2003). Using Web link analysis to detect and analyze hidden Web communities. In D. Vriens (Ed.), *Information and communications technology for competitive intelligence* (pp. 57–84). Hilliard, OH: Ideal Group.
- Rodríguez Gairín, J. M. (1997). Valorando el impacto de la información en Internet: AltaVista, el "Citation Index" de la Red. *Revista Espanola de Documentacion Cientifica*, 20, 175–181.
- Ross, N. C. M., & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science*, 51(10), 949–958.
- Rousseau, R. (1997). Sitations: An exploratory study. *Cybermetrics*, 1(1). Retrieved November 20, 2003, from <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>
- Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3(1). Retrieved November 20, 2003, from <http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html>
- Rudner, L. M., Gellmann, J. S., & Miller-Whitehead, M. (2002). Who is reading on-line education journals? Why? And what are they reading? *D-Lib Magazine*, 9(12). Retrieved November 20, 2003, from <http://www.dlib.org/dlib/december02/rudner/12rudner.html>
- Rusmevichientong, P., Pennock, D. M., Lawrence, S., & Giles, C. L. (2001). Methods for sampling pages uniformly from the Web. *Proceedings of the AAAI Fall Symposium on Using Uncertainty Within Computation*, 121–128.
- Sandstrom, P. E. (2001). Scholarly communication as a socioecological system. *Scientometrics*, 51(3), 573–605.
- Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum*, 33(1), 6–12.

- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269.
- Small, H. (1999). Visualizing science through citation mapping. *Journal of the American Society for Information Science*, 50(9), 799–812.
- Smith, A. G. (1999). A tale of two Webspaces: Comparing sites using Web impact factors. *Journal of Documentation*, 55(5), 577–592.
- Smith, A. G., & Thelwall, M. (2002). Web impact factors for Australasian universities. *Scientometrics*, 54(3), 363–380.
- Snyder, H., & Rosenbaum, H. (1999). Can search engines be used as tools for Web-link analysis? A critical view. *Journal of Documentation*, 55(4), 375–384.
- Spink, A., Jansen, B. J., & Ozmutlu, H. C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research*, 10(4), 317–328.
- Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(2), 107–109.
- Spink, A., Ozmutlu, S., Ozmutlu, H. C., & Jansen, B. J. (2002). U.S. versus European Web searching trends. *SIGIR Forum*, 32(1), 30–37.
- Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226–234.
- Steyvers, M., & Tenenbaum, J. B. (2001). *The large-scale structure of semantic networks: Statistical analyses and a model for semantic growth*. Retrieved November 20, 2003, from <http://arxiv.org/pdf/cond-mat/0110012>
- Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management*, 28(1), 1–3.
- Tang, R., & Thelwall, M. (2003). Disciplinary differences in US academic departmental Web site interlinking. *Library & Information Science Research*, 25(4), 437–458.
- Thelwall, M. (2000a). Commercial Web sites: Lost in cyberspace? *Internet Research*, 10(2), 150–159.
- Thelwall, M. (2000b). Web impact factors and search engine coverage. *Journal of Documentation*, 56(2), 185–189.
- Thelwall, M. (2000c). Who is using the .co.uk domain? Professional and media adoption of the Web. *International Journal of Information Management*, 20(6), 441–453.
- Thelwall, M. (2001a). Exploring the link structure of the Web with network diagrams. *Journal of Information Science*, 27(6), 393–402.
- Thelwall, M. (2001b). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52(13), 1157–1168.
- Thelwall, M. (2001c). The responsiveness of search engine indexes. *Cybermetrics*, 5(1). Retrieved December 9, 2003, from <http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html>
- Thelwall, M. (2001d). Results from a Web impact factor crawler. *Journal of Documentation*, 57(2), 177–191.
- Thelwall, M. (2001e). A Web crawler design for data mining. *Journal of Information Science*, 27(5), 319–325.
- Thelwall, M. (2001f). Web log file analysis: Backlinks and queries. *ASLIB Proceedings*, 53(6), 217–223.
- Thelwall, M. (2002a). A comparison of sources of links for academic Web impact factor calculations. *Journal of Documentation*, 58(1), 60–72.
- Thelwall, M. (2002b). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites. *Journal of the American Society for Information Science and Technology*, 53(12), 995–1005.

- Thelwall, M. (2002c). Evidence for the existence of geographic trends in university Web site interlinking. *Journal of Documentation*, 58(5), 563–574.
- Thelwall, M. (2002d). Methodologies for crawler based Web surveys. *Internet Research*, 12(2), 124–138.
- Thelwall, M. (2002e). A research and institutional size based model for national university Web site interlinking. *Journal of Documentation*, 58(6), 683–694.
- Thelwall, M. (2002f). Research dissemination and invocation on the Web. *Online Information Review*, 26(6), 413–420.
- Thelwall, M. (2002g). The top 100 linked pages on UK university Web sites: High inlink counts are not usually directly associated with quality scholarly content. *Journal of Information Science*, 28(6), 485–493.
- Thelwall, M. (2003a). Can Google's PageRank be used to find the most important academic Web pages? *Journal of Documentation*, 59(2), 205–217.
- Thelwall, M. (2003b). Web use and peer interconnectivity metrics for academic Web sites. *Journal of Information Science*, 29(1), 11–20.
- Thelwall, M. (2003c). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3). Retrieved November 20, 2003, from <http://informationr.net/ir/8-3/paper151.html>
- Thelwall, M. (2004). Methods for reporting on the targets of links from national systems of university Web sites. *Information Processing & Management*, 40(1), 125–144.
- Thelwall, M. (in press a). Can the Web give useful information about commercial uses of scientific research? *Online Information Review*, 28(2).
- Thelwall, M. (in press b). Data cleansing and validation for multiple site link structure analysis. In A. Scime (Ed.), *Web mining: Applications and techniques*. Hershey, PA: Idea Group.
- Thelwall, M. (in press c). *Scientific Web intelligence: Finding relationships in university Webs*. Wolverhampton, UK: University of Wolverhampton.
- Thelwall, M., & Aguillo, I. F. (2003). La salud de las Web universitarias españolas. *Revista Española de Documentación Científica*, 26(3). Retrieved January 29, 2004, from [http://www.eicstes.org/EICSTES_PDF/PAPERS/La%20salud%20de%20las%20Web%20universitarias%20espa%C3%B1olas%20\(Thelwall-Aguillo\).PDF](http://www.eicstes.org/EICSTES_PDF/PAPERS/La%20salud%20de%20las%20Web%20universitarias%20espa%C3%B1olas%20(Thelwall-Aguillo).PDF)
- Thelwall, M., Binns, R. Harries, G. Page-Kennedy, T. Price, E., & Wilkinson, D. (2002). European Union associated university Websites. *Scientometrics*, 53(1), 95–111.
- Thelwall, M., & Harries, G. (2003). The connection between the research of a university and counts of links to its Web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science and Technology*, 54(7), 594–602.
- Thelwall, M., & Harries, G. (2004a). Can personal Web pages that link to universities yield information about the wider dissemination of research? *Journal of Information Science*, 30(3), 243–256.
- Thelwall, M., & Harries, G. (2004b). Do better scholars' Web publications have significantly higher online impact? *Journal of the American Society for Information Science and Technology*, 55(2), 149–159.
- Thelwall, M., & Smith, A. (2002). A study of the interlinking between Asia-Pacific university Web sites. *Scientometrics*, 55(3), 363–376.
- Thelwall, M., & Tang, R. (2003). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics*, 58(1), 153–179.
- Thelwall, M., Tang, R., & Price, E. (2003). Linguistic patterns of academic Web use in Western Europe. *Scientometrics*, 56(3), 417–432.
- Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2), 162–176.

- Thelwall, M., Vaughan, L., Cothey, V., Li, X., & Smith, A. (2003). Which academic subjects have most online impact? A pilot study and a new classification process. *Online Information Review*, 27(5), 333–343.
- Thelwall, M., & Wilkinson, D. (2003a). Graph structure in three national academic Webs: Power laws with anomalies. *Journal of the American Society for Information Science and Technology*, 54(8), 706–712.
- Thelwall, M., & Wilkinson, D. (2003b). Three target document range metrics for university Web sites. *Journal of the American Society for Information Science and Technology*, 54(6), 489–496.
- Thelwall, M., & Wilkinson, D. (2004). Finding similar academic Web sites with links, bibliometric couplings and colinks. *Information Processing & Management*, 40(3), 515–526.
- Thomas, O., & Willett, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science*, 26(6), 421–428.
- Valverde, S., Ferrer i Cancho, R., & Solé, R. V. (2002). *Scale-free networks from optimal design*. Retrieved November 20, 2003, from <http://arxiv.org/abs/cond-mat/0204344>
- van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics*, 47(2), 347–362.
- van Raan, A. F. J. (2001). Bibliometrics and Internet: Some observations and expectations. *Scientometrics*, 50(1), 59–63.
- Vaughan, L. (in press). Web hyperlink analysis. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement*. San Diego, CA: Academic Press.
- Vaughan, L. (2004). Web hyperlinks reflect business performance: A study of U.S. and Chinese IT companies. *Canadian Journal of Information and Library Science*, 28(1), 17–31.
- Vaughan, L., & Hysen, K. (2002). Relationship between links to journal Web sites and impact factors. *Aslib Proceedings*, 54(6), 356–361.
- Vaughan, L., & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? *Journal of the American Society for Information Science and Technology*, 54(14), 1313–1324.
- Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of the American Society for Information Science and Technology*, 54(1), 29–38.
- Vaughan, L., & Thelwall, M. (in press a). A modeling approach to uncover hyperlink patterns: The case of Canadian universities. *Information Processing & Management*.
- Vaughan, L., & Thelwall, M. (in press b). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management*.
- Vaughan, L., & Wu, G. (2003). Link counts to commercial Web sites as a source of company information. *Proceedings the 9th International Conference of Scientometrics and Informetrics*, 321–329.
- Wang, P., Berry, M., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743–758.
- Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology*, 105(2), 493–527.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of “small-world” networks. *Nature*, 393, 440–442.
- Weare, C., & Lin, W. Y. (2000). Content analysis of the World Wide Web: Opportunities and challenges. *Social Science Computer Review*, 18(3), 272–292.
- webopedia.com (2003). *World Wide Web*. Retrieved November 20, 2003, from http://www.webopedia.com/TERM/W/World_Wide_Web.html

- Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science*, 29(1), 59–66.
- Wilkinson, D., Thelwall, M., & Li, X. (2003). Exploiting hyperlinks to study academic Web use. *Social Science Computer Review*, 21(3), 340–351.
- Wolfram, D., Spink, A., Jansen, B., & Saracevic, T. (2001). Vox populi: The public searching of the Web. *Journal of the American Society for Information Science and Technology*, 52(12), 1073–1074.
- Yao, Y. Y., Zhong, N., Liu, J., & Ohsuga, S. (2001). Web intelligence (WI): Research challenges and trends in the new information age. *Lecture Notes in Artificial Intelligence*, 2198, 1–17.
- Zanasi, A. (1998). Competitive intelligence through data mining public sources. *Competitive Intelligence Review*, 9(1), 44–54.
- Zhao, D., & Logan, E. (2002). Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. *Scientometrics*, 54(3), 449–472.
- Zhu, J., Hong, J., & Hughes, J. G. (2001). PageRate: Counting Web users' votes. *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, 131–132.
- Zhu, J., Hong, J., & Hughes, J. G. (2002). Using Markov models for Web site link prediction. *Proceedings of the 13th ACM Conference on Hypertext and Hypermedia*, 169–170.
- Zhu, J., Hong, J., & Hughes, J. G. (in press). PageCluster: Mining conceptual link hierarchies from Web log files for adaptive Web site navigation. *ACM Transactions on Internet Technology*.

5. Cross-spherical analysis and other forms of comparative media studies

R. Rogers and N. Marres (2002).
"French scandals on the Web and on
the street: Experiment in stretching the
limits of reported reality." *Asian Journal
of Social Science*. 30(2): 339-353.

French Scandals on the Web, and on the Streets: A Small Experiment in Stretching the Limits of Reported Reality

RICHARD ROGERS

*Department of New Media Studies
University of Amsterdam*

NOORTJE MARRES

*Department of Philosophy
University of Amsterdam*

The main argument of this paper is that the Web is well suited for the role of a streetwise informant for protest events. Using special software, we plot the network of Web sites of organisations related to the meat and anti-globalization protests in Milau, France, in June of 2000, and compare our findings about the nature of the actors with the reports in the mass media. Not only do the findings deepen mass media accounts but they also point to a new web-sociological means to re-open the question of the real. Putting on display the institutional dimension of the protest, the Web network simultaneously undermines and enriches the mass media accounts of the protests. The Web may provide a reality check on existing accounts of the protest and document encounters between opposing definitions of reality. Our analysis also suggests the Web can be harnessed to complicate the definition of real social action.

Methodological Dictum

If it's not on the streets, it may be on the Web

But if it's not on the Web, and it's not on the streets . . .

Case closed?

Farmers are on the streets, but who are these farmers?

Better check the Web!

Academic Graffiti, 2000

Introduction: Newsreel

On 30 June 2000, the Dutch daily newspaper *de Volkskrant* published a short report on the globalization protests taking place in the French town of Milau. The report referred to a “bunch of disorganized anarchists” engaged in “activist tourism”. The piece also cast doubt on the viability of this form of social criticism. Referring to the smashing of windows at

AJ.S.S. 30:2 (339-353)

© 2002 Koninklijke Brill NV, Leiden

also available online

see www.brill.nl

340 • Richard Rogers and Noortje Marres

'MacDo' as a case in point, the piece expressed concern about the corner that activism has been backing itself into ever since the violent encounters during the WTO meeting in Seattle in 1999.

Inevitably, the Internet had a part to play in these accounts. The advent of anti-globalization protest in its current form is closely tied to the wiring up of society. The Internet was introduced as an essential for the widespread dissemination of announcements for upcoming events, as well as for in situ logistical support for the protesters. What was surprising about the role allocated to the Internet in these accounts was its depiction as a medium of convenience. Just as the activists were portrayed as youthful, backpacking travellers, the Internet was characterized as their guidebook in this particular context. The Internet served as a kind of *Lonely Planet* edition.

In this paper we will argue that the Internet — or the World Wide Web — may serve as a substantive guide to the events in Milau, France. The location and analysis of the Milau protest network on the Web will show how the Web may have provided insights into the substantial organizational efforts that were behind the event. What is more, these organizational accounts potentially served as adequate means of evaluating the larger rationale of the event. Our empirical analysis of the role of the Internet in recent global protest events, thereby, potentially undermines mass media accounts of the same. But more importantly, it aims to show how the Web undermines and enriches accounts of anti-globalization protest events themselves and the Internet. In the process, questions about the part played by social science in the broader societal critique of mass media arise, as do questions about the broader societal commitment to the new medium, not to mention social science's commitment to that commitment.

The definition of the Internet as a channel of alternative reporting, which would potentially make up for the deficits of mass media, has been evoked by both mass and Web media ever since the integration of new media into publishing practices took off about eight years ago. Moreover, the promise of alternative media, historically, has a special status within social science. Recently, Steve Jones (2000:172) has pointed to the deep-rooted connection between social science and alternative reporting, citing an adage by John Dewey that "a proper daily newspaper would be the only possible social science." While Jones is quick to reject a simple transplantation of Dewey's grand aspirations onto the Web — by which the Internet would be embraced as candidate for bridging the great divide between social science and the public — he does single out the definition of the Web as an alternative to off-line media. According to Jones, it is in its guise as an alternative to established reporting practices that the Web becomes an important object of social science reporting. In the case study that follows, we take up this understanding of the Web as an alternative as an empirical question worthy of investigation.

From our controlled exercise of searching the Web for alternative accounts of mass media events, we have two key findings. On the one hand, from the vantage point of empirical study, the Web cannot be properly understood as a parallel universe, existing 'alongside' the world of mass media. On the other hand, the exercise indicates that the Web may be especially well equipped to serve as an informant of sociological accounts of events reported in the mass media.

Romanticisms of the Web and of the Streets: the Web as 'Reality Check'

The definition of the Internet as a realm of alternatives has been part of the repertoire of social studies of the Internet. The question of whether interactions via the Internet 'complement' or 'substitute' their off-line manifestations, for example, has occupied social scientists. As Sherry Turkle (1999:341) puts it, it proves difficult to keep out of the 'war of the worlds' while defending or resisting the 'virtual life'. Even the post-disciplinary framing of a research field called 'Internet studies' already contains the seeds of this conflict (Stone, 1995; Borgman, 1999). It points towards the definition of the Internet as a domain unto itself, juxtaposed with social lives off the Net. While this definition of the Internet can, of course, at all times be dismissed on methodological, theoretical, or, as we will show, empirical grounds; the simple fact of dedicating one's inquiries to the one channel in particular, brings along with it the risk of inadvertently choosing sides in the conflict between the virtual and the real, indeed, favouring the former.

We became particularly aware of this divide when asked by a French sociologist whether we, as Web analysts, could explain why the French did not hit the streets after a fraud scandal surrounding the *Crédit Lyonnais* bank. The initial suspicion handed to us was that the French may instead have 'hit the Web'. As Web analysts within the broader field of the sociology of scandals and protests, we sought not just to dig up alternative accounts, but also to check whether the Web perhaps provides an alternative to the streets. The definition of the Internet as alternative, not just to the news, but also to the off-line in general in this respect, seemed to come with the job of being the Web sociologist among sociologists.

Turning this problem into a virtue, we decided to go after French protesters on the Web, as a way of testing the understanding of the Web as a realm of alternatives. Using the earlier methodological dictum as a guideline, our intention was to provide a reality check for the idea that the Internet has the potential to somehow substitute the mass media, or for empty streets. We will, thus, test the twin assumptions that the Web provides an alternative to mass media reporting, and to the streets. As to

342 • Richard Rogers and Noortje Marres

the first assumption, our small experiment of ‘checking the Web’ provided an empirical answer to the question of whether the Web can be approached as a realm that exists alongside that of mass media reporting. As to the second assumption, the aim of our small experiment was to see if and how the Web can come to our aid in filling in, and complicating, the picture of the streets. To the degree that the Web allows for a depiction of the substantial organizational efforts that go into street protests, the small experiment also provides a reality check for the romanticism of the streets — found in the mass media as well as in some social science circles — which equates real social action with street gatherings. Our small experiment, in this way, aims to provide a test of the explanatory power of the Web; on whether it can provide an account of events that add to mass media and social science reporting.

In a lecture to the British Virtual Society Research Centre, Bruno Latour (1998), argued that the Web is mainly of importance to social science insofar as it makes possible new types of descriptions of social life. According to Latour, the social integration of the Web constitutes an event for social science because the social link becomes traceable in this medium. Thus, social relations are established in a tangible form as a material network connection. We take Latour’s claim of the tangibility of the social as a point of departure in our search for French protesters on the Web.

Encounters with the Street and the Web, and the Indispensability of ‘Webbified’ Mass Media

It is important to point out that we know little of the Crédit Lyonnais scandal from our vantage points in Amsterdam and Budapest, and from our respective wired sources of knowledge and information. We hear that the French did not hit the streets because of the Crédit Lyonnais scandal. All we know in terms of the French hitting the streets, recently, concerns not bankers but French farmers. While not wishing to make too much of our semi-ignorance, we consider the circumstance favourable in light of our question, i.e., what *the Web* — not merely TV and the newspapers — can tell us about the French protests and scandals on the streets.

In asking whether the French hit the Web as an alternative to the streets, we are of the impression that an affirmative answer would imply a radical transformation of the rules of mobilization as they are perceived by the mass media. The problem with the Crédit Lyonnais scandal, as we speculated, was likely to be a problem of iconization. The scandal’s focal points — the *faux bilans* published by Crédit Lyonnais and the ‘substantial costs per French taxpayer’ — assumed the form of numbers. Was it the scandal’s resistance to movement from this numerical realm into the more

colourful, more material realm of media icons that kept the French off the streets? If so, a solid presence of protesting French on the Web as opposed to the streets would mean that this medium in some way evaded the golden rule assumed by the mass media: iconization as indispensable for mobilization. When boarding the Web, it is precisely the hope to short cut the mass media and their golden rule of iconisation that has to be stowed. In order for the uninitiated to find protesting French on the Web, a familiar starting point is required; this can be none other than a mass search engine like www.altavista.fr or a mass 'journal' like www.tout.lemonde.fr.

For our attempt to locate a displaced protest surrounding the Crédit Lyonnais scandal, www.altavista.fr and www.tout.lemonde.fr provided the starting points. A story from www.tout.lemonde.fr yielded the names of the main actors involved in the scandal, among which were a number of potentially 'http-ed' institutions such as Crédit Lyonnais, MGM, Commission des Opérations de Bourse, la Cour des comptes, l'Inspection des Finances, le Trésor, etc. With the aid of the search engine, their URLs can be found. But these actors, even if they mention the scandal, do not link to one another. A web-based network of organizations involved in or mobilizing around the scandal could not be located. In the absence of any direct acknowledgement of other parties involved among our actors, we were forced to conclude that the Web showed no sign of collective engagement with the issue.

The question of whether the Web allows for mobilization in the absence of media iconization had to be, at least for this case and at that moment, answered in the negative. In this instance, we admit we were unable to harness the explanatory power of the Web. The only way the absence of the Crédit Lyonnais issue on the Web could provide an explanation for the empty streets would be the argument that the French did not hit the streets *because* they failed to hit the Web. We would not want to go that far. Secondly, we must admit that from the standpoint of the uninitiated, an unfulfilled romanticism of the streets cannot be replaced by a romanticism of the Web. The streets are empty, but so is the Web. In the case of empty streets, the hope that mobilization may occur through the Web must be abandoned. Finally, we gladly admit that, at least for the moment, there is no need to start grappling with the issue of the absorption of the French streets by the Web. Thus, we partially fill in the dictum: There are Frenchmen in the streets and not on the Web in the Crédit Lyonnais case. But there are farmers in the streets.

The television representation of French farmers on the streets looked like proper street protest. The newspaper articles, however, presented a less straightforward story, questioning the farmers' authenticity. [Www.tout.lemonde.fr](http://www.tout.lemonde.fr) alluded to the 'phoneyess' of the farmers; the Dutch newspaper *de Volkskrant* subsequently referred to a 'bunch of disorganised

344 • *Richard Rogers and Noortje Marres*

anarchists', joyfully hopping on the protest train, whose act of 'activist tourism' thereby compromised the original cause of the farmers. The involvement of an impressive range of international entities in the trial following the farmer attack on a 'Macdo' seems to complicate the picture further. It is hard to believe that the international entities just dropped out of the air, as the romanticist perspective that portrays street protest as a spontaneous phenomenon might have it.

We return to our dictum:

There are Frenchmen not in the streets and not on the Web (Crédit Lyonnais)
There are farmers in the streets; but who are these farmers? (Better check the Web!)

From the newspapers, we seem to have stumbled upon the delicate subject of the French streets potentially being invaded, perhaps even taken over, by international entities. Because it is not just the romanticism of the streets, but of French streets that seems shaken, the issue of the identity of the protesters becomes vital. How does one find out? Better check the Web! Whilst it is the case that, for the uninitiated, protests on the Web cannot be seen as distinct from protests on the streets, from real worldly icons; the Web may nevertheless provide us with the means to stretch the limits of iconization and gain a clear view of what the farmers do besides farming.

Web Findings: Who are these Farmers? Whose are these Streets?

Following the same methodology as outlined above, involved actors were distilled from an article at www.tout.lemonde.fr, and their URLs easily found. Interestingly, even though the fact cannot be perfectly squared with allusions to 'phoney farmers' made elsewhere on www.tout.lemonde.fr, our finding indicates French-only starting points, with the semi-exception of Attac, the NGO founded in France, and now internationalized (see Figure 1). Our list of starting points is in keeping with the historicized French street romance of a spontaneous alliance of students and workers. The impression is only reinforced by that fact that quite a few of the websites of the syndicates, political party branches and associations mentioned by the newspaper, do mention the farmers' trial. Was the gathering in the town of Milau, after all, predominantly French, and maybe even ad hoc?

We now turn to our homemade 'netlocator' software to find out whether an authoritative network is disclosed from our chosen starting points. If it locates such a network, the 'netlocator' would give us an indication of the degree to which the streets are still romantic, still French. With the netlocator serving as our medium, the Web was searched in an

Figure 1. Actors in “French Farmers Protest” Media Story



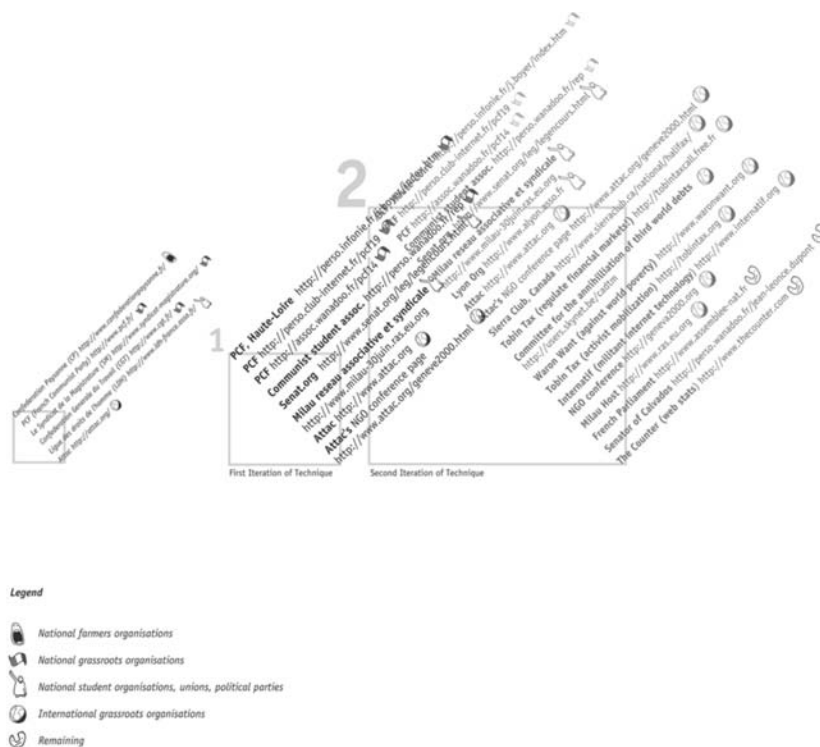
attempt to re-make or unmake the various spectacles brought to us by CNN, *de Volkskrant* and www.tout.lemonde.fr.

Actors from the newspaper article that (a) have a website, (b) mention the farmers' trial and (c) link to other domains, were entered into the netlocator. Links from the URLs of the French farmers, French left wing political parties and syndicates, French NGOs, and a network were returned through co-link analysis. This network contained mainly French political parties, syndicates and NGOs. The farmers are absent. The brief analysis yielded the view that, according to the Web, the farmers were not farmers, but mainly French 'politicos'. Without the farmers, we had only a portion of the ingredients for the romantic streets recipe. Taking these French politicians and the initial starting points, and then inputting them into the netlocator, however, returned not only French politicians, but also a significant number of international, issue-based, activist organizations, many of which are dedicated to global economy issues. There were still no farmers. Thus, according to the Web, the farmers were not farmers, but represented an organizational configuration that moved from the national to the global, and from the political-ideological to the issue-activist. As is evident from Figure 2, it was quite an organized picture, whereby not farmers, nor 'phoney farmers', nor 'a bunch of disorganized anarchists' made up the protests. Instead, it was a professional national-international network.

It is important to stress that merely querying the Web does not allow the uninitiated to locate the protest network. We were unable to evade the media's narrow definition of what counts as real protest. But the Web enabled us to put the media's rehearsals of the events as well as our friendly analysts' romanticism of the streets in their proper places, as it shows us that, like Russian dolls, organized professionals inhabited the iconized and perhaps romanticized figures of farmers (see Figure 3). By means of the Web, the limits of iconization and street romanticism can be stretched.

346 • Richard Rogers and Noortje Marres

Figure 2. Actors in “French Farmers Protest” Network on the Web.



Here, we would like to make clear, to the proverbial French romantic, that the Web does not remedy the dying romanticism of the streets. Rather, the streets are alive with ‘webby’ networks.

Conclusions

We wish to conclude with the idea that the main virtue of the virtual is to open up the question of the real. There are four steps in this position. Firstly, one could be tempted, from the outset, to believe in the purity of French farmer protest and the streets, especially if one is a foreigner. Secondly, upon examination of the media, the reality of the event of French farmer street protest is compromised when it is cynically represented and mystified as ‘a bunch of disorganized anarchists’, engaged in a novel form of conflict tourism. Such a rendering also unintentionally blurs journalism with cultural studies. Thirdly, ‘checking the Web’ becomes a ‘reality check’ in the sense that it allows one to fill in the integrity of the streets that were sullied and compromised through reporting. Of course, one could

Figure 3. "From the Web to the streets: French Farmers as Russian Dolls".



travel to Milau, and by observing and/or participating, capture some of the complexity of the French streets in the age of the Web. But the streets tend to have no names. Indeed, reports from Seattle often lacked any semblance of knowledge about the actual networks of groups involved and their positions. Sound bites on TV and one-liners in newspaper articles often strip "a group calling itself..." of a (networked) past and a (networked) future. Only an overall anti-message from a group of protestors, plus TV spectacle, was communicated. Fourthly, and finally, it could be argued that the Web and Web analyses ultimately enriches the streets. Without it, the 'coded' website for 'swimmers' announcing team 'meets', as was in the case of Milau, France on 30 of June, would remain opaque.

In adding the description of the organizational network that supported it to the mass media accounts of the street protest, our small exercise brought in the web and a research instrument to take up the formidable task put to social science by C. Wright Mills (1971:212) as "no less than to present conflicting definitions of reality itself." Whereas the goal of Mills' social science was to extract the marginal definitions of reality from social life by means of the interview and the survey, Web analyses can rely on pre-existing documentations of social life, having been encoded on to the Web by the actors 'themselves'. In the context of the new media, the staging of the conflict between competing definitions of realities thus becomes a question of crossing available information streams, to bring those emanating from news agencies in contact with those generated on the Web. Insofar as the justification for this type of research is concerned, however, Mills' mission statement can be transplanted to the context of the new media

348 • Richard Rogers and Noortje Marres

virtually unchanged. It was Mills' conviction that as documenter of the conflict between dominant and marginal definitions of reality, social research counts as a viable alternative to "hitting the pavement, taking the next plane to the scene of the current crisis [. . .] or buying a newspaper plant." Web analysis adds descriptions of the encounter between action and reporting. But as a footnote to Mills, we would like to add that in documenting that encounter on the Web, it is possible to show that *political action itself* may extend well beyond the narrow limits to action suggested by 'hitting the pavement' or 'buying the plant'.

Methodological Appendix

We ask, who are these farmers? Or, how are we able to find (or be sure not to find) French farmers on the Web, and determine whether *they* are hitting the streets? To make a determination, one first finds the 'issue network' — those (1) discussing and debating the 'issue' and (2) mobilizing other actors to action. In order to find an issue network among the swells of sites whose issues somehow revolve around protesting French farmers, one is in need of fixed starting points. Previously, we have identified and discussed the outcomes of at least five distinct starting points for locating issue networks on the Web: search engines, associative reasoning, media stories, public actors and/or discussion lists (Rogers/Zelman, 2001). In isolation or in combination, these means of identifying starting points are aimed at finding those organizations whose sites reveal the debate most extensively by virtue of their respective link lists. This is the first step in locating an issue network. We then deploy a simple, home-made machine (the "netlocator" also called the 'depluralising engine' by www.govcom.org) to 'rub' the network and chart the most relevant sites. By 'rub', we mean that the netlocator mines each starting point three levels deep, follows links and identifies candidate sources; those sources linked by at least two actors (in the 'medium inclusiveness' setting) are brought back by the locator. The netlocator currently supports up to seven starting points. Normally, a minimum of three will suffice.

As for the different means of locating the starting points with search engines, it is assumed that the keywords are known by the surfer. One uses the top returns as starting points, and rubs for a network with the machine. Associative reasoning relies on intelligent guesswork; either the issue or the presumed relevant organization is simply typed into the browser, with .org, .com or another suffix attached (in a previous case study, we began with milk.org, grains.org and corn.org). One follows links from the associatively reasoned site(s), until one finds sites displaying the debate extensively. These sites are rubbed. The third technique allows a leading media story to be the guide; those organizations mentioned are then located either through a search engine, or by associative reasoning. The URLs are

then rubbed. Public actors are similarly located; one presumes well-known public actors (e.g., Greenpeace for climate change) will display the debate around an issue extensively, and they are located through a search engine or by associative reasoning. One could also subscribe to a discussion list about the topic in question, and chart the links recommended by the discussants in their ongoing postings over a particular period in time. It is important to note that each means relies on distinctly different 'expertise' or 'recommenders' with varying epistemological and 'info-societal' consequences for each of the subsequent networks located (Marres/Rogers, 2000).

We present the case study with the aid of the 'diary of a crawler'. The kept log explains how the network is located; it also details the most significant finding touched upon above, that is, *the de-iconization of the romantic French streets by organized, virtual, global civil society*. We present the case through the following steps below:

1. Knowing nothing of the issue concerning French farmers, we turned to www.tout.lemonde.fr. There was no need to take recourse to their *moteur de recherche*; the first headline on their homepage staged our farmers. The issue according to *le toutlemonde* (<http://www.lemonde.fr/article/0,2320,seq-2030-74010-QUO,00.html>):

Événement planétaire à Millau, Aveyron: 30 000 à 50 000 manifestants venus du monde entier sont attendus pour le procès, le 30 juin, de José Bové. Figure emblématique de la résistance à la mondialisation, il sera jugé pour avoir attaqué, en août 1999, le McDonald's en construction dans cette ville. La chaîne américaine CNN a installé ses caméras dans plusieurs appartements face au tribunal.

Then some background:

Le 12 août 1999, un groupe de trois cents éleveurs du Syndicat des producteurs de lait de brebis et de la Confédération paysanne «démontent» virilement le restaurant McDonald's en construction dans la ville. Avec M. Bové à leur tête, ils entendent protester contre la surtaxation américaine du fromage de roquefort après que l'Union européenne eut décidé de ne plus importer de viande aux hormones des Etats-Unis.

2. The article yielded a long list of potentially webby (http-ed) actors:

- CNN (present at the scene)
- Jose Bove (leader of the resistance)
- *le Syndicat des producteurs de lait de brebis* (farmers)
- *la Confédération paysanne* (farmers)
- Lori Wallace of Public Citizen ('expert' witness in the trial)
- *Confédération paysanne* (the farmers)
- CGT (*Confédération Generale du Travail*)

350 • Richard Rogers and Noortje Marres

- *les SUD* (*solidaire, unitaire et démocratique, a syndicat*)
 - CNT (*confederation nationale de travail*)
 - PS (Socialist party)
 - PCF (Communist party)
 - Verts (Green party)
 - LCR (*ligue communiste revolutionnaire*)
 - *Ligue des droits de l'homme* (human rights organization)
 - DAL (*droit au logement*)
 - Attac (association of citizens, newspapers and syndicats)
 - *Droits devant!* (human rights organization, linked to DAL)
 - *Le Syndicat de la Magistrature* (syndicat)
3. Most of these actors were easily found by surfing from search engine returns (www.altavista.fr) and by following links. That is to say, we were on familiar ground where the behaviour of civil society actors was concerned; we had linking organizations that disclosed their cohorts. A selection of these actors (those that had a website, discussed the issue, and linked to other actors) was ready to be fed into the network locator. Judging from the intensity of linking, a demarcated sample would have probably emerged, and we would have had the makings of an issue network map.
 4. As is often the case with freshly erupted events, two actors presented as central players by the newspaper *Syndicat des producteurs de lait de brebis* and the French McDonald's, had a very thin presence on the Web. www.altavista.fr returned only five entries when queried for the syndicat: three online newspaper articles (*midilibre*), the *confederation paysanne*, and *millau-clic*, a local portal for the town that was our scene of action. (For reasons that are easy to guess, that site was almost impossible to reach on 30 June 2000 — a protest day). The syndicat did not own its own domain on the Web, and www.Macdonalds.fr was 'under construction'. For the sake of contrast, we would like to mention that during the Kyoto Conference on Climate Change (1998), the website of Shell, a main target of the CO₂ emission reduction lobby, initially remained more or less silent on the issue. When British eco-terrorists destroyed GM crops in the summer of 1999, the Monsanto website initially showed no sign of awareness of the protests against its business. While McDonalds is certainly less central to the farmers protest, and the fact that its site was under construction most probably due to other reasons, it is surprising to notice how actors that are at the centre of the action according to the mass media, remain in the background on the Web, at least initially.
 5. Judging from its 'links in from the issue-network' the following site has a solid presence: <http://www.millau-30juin.ras.eu.org>. It is a practical information site set up by, well, some of the people involved. Self-

evidently this site did not figure in the newspaper article. Newspapers only go so far in contributing to mobilization.

6. As to the question whether 'they' were really farmers, the answer seemed to depend on the centrality of the *confederation paysanne* (farmers' federation) in the network. In any case, it was clear that there were many people in Milau dressed up as farmers. Or more accurately, the farmers were serving as 'dress' for many other players. Not surprisingly, it turned out that for .orgs, it was more a question of going to where the action was, than whether one was already tied to the issue. The organizations involved range from 'against homelessness' to 'the workers' to 'the environment'. That is to say, they did not exactly have careers in farming. They were civil society swarms.

On a final note, it should be mentioned that two domains that appeared in the Credit Lyonnais scandal also figured at <http://www.verts.imagnet.fr> and [wanadoo.fr](http://www.wanadoo.fr).

Actors in 'French Farmers Protest': Media Story Network

• <i>Confederation Paysanne</i> (CP)	http://www.confederationpaysanne.fr/
• <i>Ligue des droits de l'homme</i> (LDH)	http://www.ldh-france.asso.fr/
• Attac	http://attac.org/
• <i>CNT energie</i> (CNT)	http://assoc.wanadoo.fr/energie/
• PCF (French Communist Party)	http://www.pcf.fr/
• <i>Les Verts</i>	http://www.verts.imagnet.fr/
• <i>Le Syndicat de la Magistrature</i> (SM)	http://www.syndicat-magistrature.org/
• <i>Confederation Generale du Travail</i> (CGT)	http://www.cgt.fr/
• LCR	http://www.lcr-rouge.org/
• DAL	http://www.easynet.fr/appelsan/dal.html
• <i>Droits Devant</i> (DD)	http://www.easynet.fr/appelsan/mani.html

Having found these linking actors with so little effort, the presence of an issue-network was to be expected. Thus, the network locator was called in to crawl and cull the network.

The link lists of CP, CNT, PCF, verts and CGT above were inputted as starting points. Hit and crawl from CP, CNT, PCF, verts, CGT brought up the following actors (By hit and crawl is meant that the locator crawls the site three levels deep and follows the outward links to sites other than its own; those sites linked by at least two actors in the sample are returned by the locator).

• PCF, Haute-Loire	http://perso.infonie.fr/j.boyer/index.htm
• PCF	http://perso.club-internet.fr/pcf19
• PCF	http://assoc.wanadoo.fr/pcf14
• Communist student association	http://perso.wanadoo.fr/rep
• Attac, Toulon	http://perso.infonie.fr/lvaise/

352 • Richard Rogers and Noortje Marres

- CGT <http://www.multimania.com/cgtforclum>
- CGT <http://assoc.wanadoo.fr/ufcmpx>

It is clear that these starting points disclose a national kinship network of allied organizations. Perhaps, interestingly, *confederation paysanne* — the organization presented as central by the newspapers and one that could be regarded as such on the basis of its site — was not in the webby issue network. Also, note the frequency of the perso (personal) domain. This brings to the surface that there are people (identifiable people no less) behind this protest event! In the Crédit Lyonnais scandal, people were not self-evident. The former crawl being a trial, six starting points were next entered: CP, LDH, Attac, PCF, SM, CGT (leaving out those public actors that did not discuss the issue). Hit and crawl brought us the following:

- Attac's NGO conference page <http://www.attac.org/geneve2000.html>
- *Le Monde Diplomatique* <http://www.monde-diplomatique.fr/md/1997/12/ramonet/9665.html>
- PCF, Haute-Loire <http://perso.infonie.fr/j.boyer/index.htm>
- PCF <http://perso.club-internet.fr/pcf19>
- PCF <http://assoc.wanadoo.fr/pcf14>
- Communist student association <http://perso.wanadoo.fr/rep>
- Senat.org <http://www.senat.org/leg/legencours.html>
- *Milau reseau associative et syndicale* <http://www.milau-30juin.ras.eu.org>

Then we inputted this actor network and the original starting points. This rub would affirm the presence of an issue-network if the sites returned here overlapped with original starting points, and these actors themselves. The second rub would yield medium threshold yields:

- Attac's NGO conference page <http://www.attac.org/geneve2000.html>
- Attac Netherlands <http://www.attac.nl>
- *Le Monde Diplomatique* <http://www.monde-diplomatique.fr/md/1997/12/ramonet/9665.html>
- PCF, Haute-Loire <http://perso.infonie.fr/j.boyer/index.htm>
- PCF <http://perso.club-internet.fr/pcf19>
- PCF <http://assoc.wanadoo.fr/pcf14>
- Communist student association <http://perso.wanadoo.fr/rep>
- Senat.org <http://www.senat.org/leg/legencours.html>
- *Milau reseau associative et syndicale* <http://www.milau-30juin.ras.eu.org>
- French Parliament <http://www.assemblee-nat.fr>
- Senator of Calvados <http://perso.wanadoo.fr/jean-leonce.dupont>
- Sierra Club, Canada <http://www.sierraclub.ca/national/halifax/>
- The Counter (web stats) <http://www.thecounter.com>
- *Tobin Tax* (regulate financial markets!) <http://tobintaxcall.free.fr>

- Committee on annihilation of Third World debt <http://users.skynet.be/cadtm>
- *Waron Want* (against world poverty) <http://www.waronwant.org>
- *Tobin Tax* (activist mobilization) <http://tobintax.org>
- *Internatif* (militant internet technology) <http://www.internatif.org>
- Lyon organization <http://www.alyon.asso.fr>
- NGO conference <http://geneva2000.org>
- Milau host <http://www.ras.eu.org>

We noticed the globalizing tendencies of issue-networking. In the second rub, many more international organizations were returned, and again the *confederation paysanne* is absent. What is more, the delocalized issue of the regulation of markets re-emerged at the second rub, where it seemed to have been absorbed by French revolutionary politics after the first rub.

Acknowledgements

The authors wish to acknowledge Marieke van Dijk for designing the figures.

References

- Borgman, Albert (1999) *Holding on to Reality: The Nature of Information at the Turn of the Millennium*. Chicago: University of Chicago Press.
- Jones, Steve (2000) "The Bias of the Web", in Andrew Herman and Thomas Swiss eds., *Cultural Theory and the World Wide Web*. New York: Routledge.
- Latour, Bruno (1998) "Thought Experiments in Social Science: From the Social Contract to Virtual Society", 1st Virtual Society Annual Public Lecture, Brunel University, London, 1 April.
- Marres, Noortje and Richard Rogers (1998) "Depluralsing the Web, Repluralsing Public Debate: The Case of the GM Food Debate on the Web", in Rogers, Richard, ed., *Preferred Placement: Knowledge Politics on the Web*. Maastricht: Jan van Eyck Editions.
- Mills, C. Wright (1971) *The Sociological Imagination*. Harmondsworth: Penguin.
- Priest, Susanna (2000) *A Grain of Truth: The Media, the Public and Biotechnology*. Lanham, MD: Rowman & Littlefield.
- Rogers, Richard and Andrés Zelman (2001) "Surfing for Knowledge in the Information Society", in Elmer, Greg, ed., *Critical Perspectives on the Internet*. Lanham, MD: Rowman & Littlefield.
- Stone, Allucquere Rosanne (1995) *The War of Desire and Technology at the Close of the Mechanical Age*. Cambridge, MA: MIT Press.
- Turkle, Sherry (1999) "Commodity and Community in Personal Computing", in Schön, Donald et al. eds., *High Technology and Low-Income Communities*. Cambridge, MA: MIT Press.

Richard W. Wiggins (2001). "The Effects of September 11 on the Leading Search Engine." *First Monday*, 7(10), <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/890/799> (accessed 18 October 2010).



The Effects of September 11 on the Leading Search Engine

by Richard W. Wiggins



The apparent terrorist attacks of September 11, 2001 changed the skyline of New York City, and the political and emotional landscape of the United States. The attack may have also changed how the leading search engine, Google, thinks of itself. This article examines how people used the Internet in general, and Google in particular, to seek and to deliver desperately wanted information about the lives lost and damage inflicted by the attacks.

Contents

[Introduction](#)
[What Users Expected from Google](#)
[Google on September 11: To Browse, or To Search?](#)
[Google Evolves After the Disaster](#)
[Ultimately, Google Acts Like a Portal](#)
[Google Use in Context](#)
[Google Responds](#)
[Where Does Google Go from Here?](#)
[Footnote: A Sadly Ironic Death](#)

Introduction

The September 11 attacks on the United States caused millions of people to urgently seek information about what had happened, who had been killed, what damage had been inflicted, and what new developments were taking place. The public at large wanted to learn all they could about attacks on their homeland. A smaller group -- those who had friends, family, or colleagues who might have been in the World Trade Center, at the Pentagon, or in one of the hijacked airliners -- sought information with even greater urgency.

When breaking news occurs, different people flock to different news sources. Many find a television and tune to CNN (or one of its growing rivals). Some turn to National Public Radio. Some open a Web browser and point it at a trusted newspaper site, such as nytimes.com or washingtonpost.com -- or the online version of a media outlet, such as cnn.com or msnbc.com.

Still others head to their favorite Web portal. Many will point browsers at Yahoo or AltaVista or Lycos or Excite. Most of these portals began their existence as robotic Web indexers. In the early days, these "search engines" in the classic sense of the term were poor places to find news of breaking events. Their crawling and indexing processes paid no special attention to news sources, and they had no mechanism to turn the spider's attention to relevant news sources. The day Princess Diana died, I spent several hours searching the then-popular search engines to see which one would first report her death. (The "winner" was Infoseek.)

Since that time, every popular search engine morphed into a portal. Human-edited content filled the screen. News links, product links, advertisements, and every manner of special service, from translation to image searches, cluttered each new portal's page. But the overwhelming dominance of human-edited Yahoo, and the inability of the search engines to effectively deliver instantaneous access to breaking news, fueled the transformation from pure engine to portal.

One search engine, relative newcomer Google, has resisted the temptation to become an all-encompassing, one-stop-shopping, complex portal. Before the September 11 attacks, Google only offered the sparest of home pages -- a simple logo, a single search box, and a handful of links to Google's own services. Not even a single banner ad appeared on the Google home page.



**Google home page
the morning of the
attack.**

Click on thumbnail to
see full-size image.

Google resisted the temptation to become yet another search engine-turned-portal. The only event-related content one ever noticed on Google was its amusing and tastefully crafted subtle changes to its logo, in order to honor holidays. (See <http://www.google.com/holidaylogos.html>.) Prior to September 11, Google devoted itself to a vision of "search is king." The user is tempted to do one thing: type a search phrase into the single form on screen, and hit Search. And the evidence is that millions of users did just that, beginning just moments after the attacks.

What Users Expected from Google

Google offers a window into what people are looking for week-by-week at <http://www.google.com/press/zeitgeist.html>. After the September 11 tragedy, Google offered a special page detailing changes in search patterns; see <http://www.google.com/press/zeitgeist/9-11-search.html>. The evolving patterns are fascinating. Note, for instance, Google's report of the top 10 search terms related to the tragedy used in searches the week after the tragedy:

Top 10 Google Queries Related to Terrorist Attacks - 9/11/01

1. [cnn](#)
2. [world trade center](#)
3. [bbc](#)
4. [pentagon](#)
5. [msnbc](#)
6. [osama bin laden](#)
7. [nostradamus](#)
8. [american airlines](#)
9. [fbi](#)
10. [barbara olson](#)

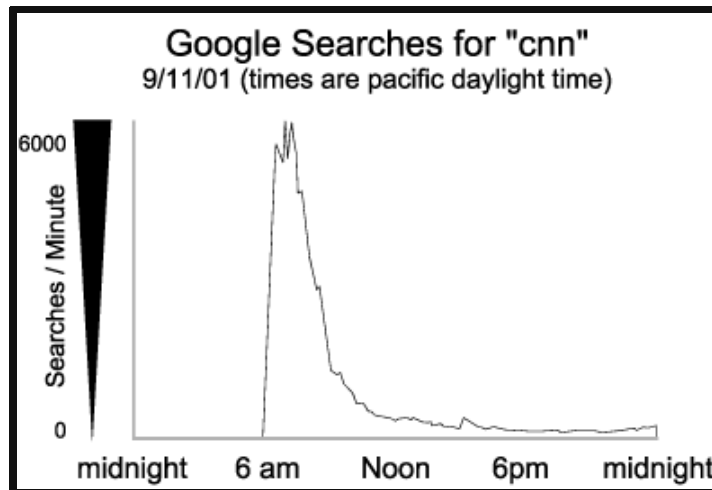
Overall, Google reports that on September 11 searching for news-related content increased by a factor of 60. For the most part, the entries are not surprising. One can imagine what various users sought:

- **world trade center:** People wanted to know more about the trade center complex: its size, location, height, architectural history, structural makeup, surroundings, occupants, etc.
- **pentagon:** People sought the same kind of information about the Pentagon, as well as perhaps any official U.S. government information about the building or the unfolding disaster.
- **osama bin laden:** Bin Laden was named a "prime suspect" soon after the hijackers were identified.
- **nostradamus:** Rumors quickly spread that Nostradamus had predicted the disaster. People no doubt sought his original words and any interpretation as to how they might have applied in this modern event.
- **american airlines:** Since two of the hijacked flights belonged to American Airlines, people sought information about the airline, and perhaps sought its official Web site (which is not obvious because its domain is "aa.com".)
- **fbi:** Obviously the Federal Bureau of Investigation plays a key role in the ongoing search for links to the hijackers.
- **barbara olson:** This notorious conservative commentator was on board American Airlines flight 77, which crashed into the Pentagon; she called her husband, U.S. Solicitor General Theodore Olson, during the hijacking to seek counsel.

The other items in the top 10 are somewhat surprising. Consider, at the top of the list, "cnn". Immediately after the attack, some 6000 users per minute used Google to find CNN. Ever since the Gulf War, millions of Americans turn to CNN instinctively in time of crisis or major breaking news. It is interesting to see that that instinct translates to online quests for breaking news. But why would a user point his or her Web browser at Google.com, and do a search engine lookup, in order to find the Web site of the Cable News Network? Why wouldn't the user simply enter "cnn" or "cnn.com" into the Address field of their Web browser?

In some cases, it's possible that the user did try to go to [cnn.com](#), and, encountering a timeout due to server or link congestion, tried Google to find the "correct" address. It seems more likely that Google is so effective at delivering popular sites at the top of the hit list that users trust Google as the primary locator for highly-popular sites. Many users have Google as their default home page; it is easier and more effective to hit the "Home" button and type the name of the site you are looking for in the Google search box.

Savvy Web surfers might consider someone naive for using a search engine to locate CNN on the Web. But if we count keystrokes and mouse clicks, it may be that this behavior is not so naive after all. Suppose a user tries to type in the URL for CNN into the browser Address field. They may type `www.cnn.com` or the proper form of `http://www.cnn.com`. Assuming Google is the user's default home page, the user need only type `cnn` into the search box, hit Search, and CNN will be the first site on the hit list. Entering in the full address costs 12 to 18 keystrokes and a mouse click. The Google approach costs 3 keystrokes and 2 mouse clicks. This user's application of Google as a finding aid is not only not naive -- it is quite efficient. Indeed, even if the user has CNN bookmarked (and is sitting at the computer the houses his or her bookmarks) the Google approach is faster than scrolling through a long list of favorite sites.



The user may realize that it's possible in most modern browsers to enter `cnn` into the Address field and hit Enter. But the behavior browsers exhibit when you do this vary with the browser, the browser version, and with preference settings. Sometimes the browser will try `www.cnn.com` automatically. Other browsers may search RealNames. Others may initiate an MSN search. If instead the user goes

straight to the Google search box, behavior is predictable: they will see a hit list, and the starting point they seek is usually the first item on the hit list. (If it isn't the very first item, for popular sites, the correct starting point is almost always in the first 10 hits.) Once again, the "naive" user exhibits a rational, efficient strategy for finding the site desired. And, as with American Airlines and `aa.com`, the greater the difference between brand name and domain name, the more rational the user's strategy becomes.

If this interpretation is correct, it speaks volumes about Google's primacy as an Internet resource locator. Those of us who have observed Web resource discovery tools unfold over the last decade generally argued that human-edited finding aids such as Yahoo would be the primary choice of people looking for Internet starting points, and that Web search engines would be used when people sought more specific information found deeper within Web sites.

No doubt Yahoo remains important as a finding aid; its reported traffic and continued relatively strong financial position attest to that. But when millions of people use Google to find CNN and MSNBC, it becomes even clearer that Google is a portal. Yet it is a portal with a difference. Upon visiting the home page, the user types in the name of the site, rather than drilling down a hierarchy of categorized links. Google delivers because simple but pioneering use of link analysis guarantees that popular sites bubble to the top of the hit list. (Google does offer a hierarchical drill-down list of Web sites based on the Open Directory project. I claim that the majority of Google users use the search box, whether they are searching for specific information or media company home pages; they'd visit Yahoo if they wanted to browse a drill-down hierarchy.)

This interpretation is consistent with recent reports that for most users, the Web isn't a place for random surfing; most users re-visit a handful of their favorite sites as they use the Web each day. Google may be more efficient than searching one's own bookmarks for popular sites; it certainly is more efficient if one is away from the PC that houses one's bookmarks. Under this interpretation, brand name is at least as important as Web site domain name. Whether the user seeks CNN or American Airlines, if it's the first item on the hit list (as will be the case with Google) the user is happy.

The Google Zeitgeist from September 24 reflects changes in the focus of the public (or at least that portion of the public that searches for information using Google):

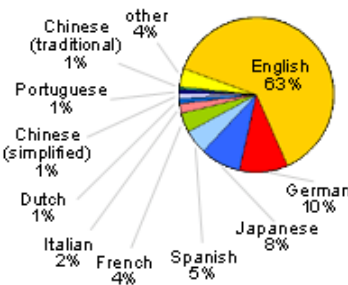
Top 10 Gaining Google Queries Week Ending Sept. 24, 2001

1. [nimda](#)
2. [counter strike](#)
3. [anthrax](#)
4. [tribute to heroes](#)

5. [miss america](#)
6. [infinite justice](#)
7. [david letterman](#)
8. [bill maher](#)
9. [friends](#)
10. [bush speech](#)

Note that this is Google's list of the top 10 *gaining* queries among *all* queries, not just those with an obvious relationship to September 11. Of course, it is easy to see that most of the items on the list bear some relationship to the tragedy. For instance, the two talk show hosts made news with their programming, Letterman for his emotional support for New York City and its police, firefighters, and mayor; Maher for his insensitive remarks about American cowardice. "Infinite Justice" was the Bush Administration's ill-chosen initial choice for a moniker for the new war on terrorism. The Nimda worm, which was released in the week after the tragedy and which may or may not have been related to the tragedy, topped the list.

One assumes that a large percentage of Google users reside in the United States, but we do not have good statistics on this. Google did report that 63% of its users in August 2001 chose to use the site in English. While assuming that the majority of U.S. residents surf the Web using English-language interfaces, there are no doubt many Google users in other English-speaking countries. (Unlike its rivals, Google does not offer country-specific portals; it does allow users to specify the language used for the Google interface, and it can limit Web searches to pages encoded in specific languages.) Given the large number of Internet users outside the United States who use English for navigating the Web, perhaps only one-half of Google's users in a typical month are American. (The presence of Letterman and Maher, known primarily by U.S. television viewers, on the "gaining" list tends to indicate that American usage dominates.) It will be interesting to see if the September figures show a different breakdown in languages. In any event, no doubt the phenomenon of using Google to find information about September 11 extends far beyond U.S. shores.



Google on September 11: To Browse, or To Search?

As we've seen, searching Google could be extremely effective for certain searches, such as looking for the CNN home page or background information on Osama bin Laden. But to the extent Google users sought real-time breaking news, a Google search is a poor place to start; Google does not currently offer a specialized index of current news sources. Moreover, many may have sought specific, authoritative sites to visit, not news per se: where to look for names of victims, Web sites for donations etc. Such users especially needed *authentic* pointers to trusted sites.

Post-September 11 demand by millions of users for real-time news and trusted Web sites may have transformed Google's vision about its role and responsibilities. Soon after news of the disaster unfolded, Google added a message to its home page informing visitors that television or radio, not the Web, was the best place to find breaking news information. Google also offered links to the Washington Post, a cached version of the CNN site, and to Yahoo News:

If you are looking for news, you will find the most current information on TV or radio.
Many online news services are not available, because of extremely high demand.
Below are links to news sites, including cached copies as they appeared earlier today.

Breaking news: Attacks hit US [Washington Post](#) - [CNN.com \(cached\)](#) - [Yahoo! News](#)

Google's advice to use a traditional broadcast medium was wise counsel. The Internet remains a narrowcast medium, not suited to millions of people flocking to a handful of media sites. People who are savvy about how the Internet works don't even try to find breaking news on the Net. I work at a major university with 1.2 gigabits/second of connectivity to the greater Internet; most of my colleagues flocked to watch cable TV in the break room, knowing that news sites would be swamped with visitors and unable to respond.

But millions of others do flock to news sites, and to portals -- and to Google. Any why not? Google is extremely effective at delivering relevant results for most of the kinds of searches real users perform. Without inside knowledge of how search engines work, doesn't it make sense to expect Google to come through with urgently-sought information when a national tragedy occurs?

Obviously the decision makers at Google knew that millions of people would rush to their site, either for breaking news, or for background information about the World Trade Center, the Pentagon, hacking, terrorism, etc. If Google officers didn't realize this, real-time monitoring of queries told them very quickly that the pattern of searching for disaster-related content had emerged. As the days unfolded after the attack, users' thirst for information shifted -- and

Google adapted its content to match the changing requests of users.

Google Evolves After the Disaster

Even though Google's advice to turn on the television was wise, Google belongs to the new media, not to the old. Eventually, Google changed their minds. Rather than sending visitors offline for current information on the unfolding tragedy, they sought to give site visitors what they were looking for.

By late afternoon Tuesday, Google's home page added to the list of news sources it was directly linking, as well as sources Google had cached:

Breaking news: Attacks hit US

Many online news services are not available because of high demand.
Below are links to news sites, including cached copies as they appeared earlier.

Current: [Washington Post](#) - [Yahoo! News](#) - [CNN](#) - [ABC News](#) - [Yahoo! News Photos](#) - [NY Times](#)

Earlier: [CNN.com](#) - [NY Times \(1\)](#) - [NY Times \(2\)](#) - [Washington Post](#)

Google's caching of content raises interesting questions. Did they have permission from CNN, The New York Times, and the Washington Post to cache their content? Even during an interval when media companies' servers and network links are strained, does Google have the right to intercept traffic that otherwise would be destined for media companies' own sites?

Offering cached versions of "earlier" content also raises questions as to how well the citizenry is being informed. When someone picks up a daily newspaper, it is obvious that the content is a snapshot of time, current as of the time the paper went to bed. The difference between "current" and "earlier" on a Web site isn't so obvious. Is "earlier" five minutes earlier? Ten minutes? An hour? Six hours?

On Wednesday, Google's home page added links to United and American Airlines -- the two airlines whose planes were hijacked and used in the attacks:

Updated news on attacks in US

Because some news sites may be unavailable, Google offers these links:

Current News: [Washington Post](#) - [Yahoo! News](#) - [CNN](#) - [ABC News](#) - [NY Times](#) - [CNET](#) - [MSNBC](#) - [BBC](#) - [more...](#)

Earlier Reports: [CNN](#) - [NY Times](#) - [Washington Post](#) - [American Airlines](#) - [United Airlines](#)

Ultimately, Google Acts Like a Portal

By late Wednesday, Google's handling of the "Breaking News" had changed yet again. Now, joining the practice of many commercial Web sites (even those with no connection to the incident or to the media), Google offered condolences to the victims. Google moved its news sources links from the home page to a new page of news and support resources.

The "new links and support information" page was far more extensive than the handful of links previously offered. Links to foreign news sources, including Arab, Israeli, and other national newspapers, were added. About this time, news broke that fraudulent donation Web sites were extracting money from an unsuspecting and malleable public. Later, we would learn that pranksters were populating some of the self-reporting "I'm okay" sites with bogus information. By offering an authoritative, human-edited list of news, donation, and rescue sites, Google provided the services one would expect of a portal. Even if Google's crawler had discovered these sites, a robotic search engine could never match the ability of a human edited list of relevant sites to separate the wheat from the chaff.

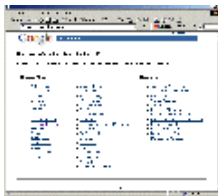
Note the links to a wide variety of news sources in the United States and abroad. Support resources include the Red Cross, the affected airlines, Web sites where survivors could identify themselves to friends and relatives, and trusted sites for giving donations. Cached links to media sites were removed; we do not know if Digital Ventures for Internet Research



Google home page late Wednesday -- the day after the attack.

Click on thumbnail to see full-size image.

content owners objected or because the overload on servers had diminished.



Google /news page late Wednesday -- the day after the attack.

Click on thumbnail to see full-size image.

Now the transformation was nearly complete. Google moved from actually *advising its visitors to leave the Web in favor of traditional broadcast media*, to providing visitors with a set of links in a variety of topical areas. In effect, Google launched a mini-portal on the disaster. Perhaps tellingly, they named the folder `google.com/news/`. Here, in Microsoft Web Archive format, is [a copy of the Google News links page as it existed on the Friday after the tragedy](#).

Google's transformation reached its ultimate form when their graphic artist created a simple logo representing the red, white, and blue ribbons many displayed after the tragedy. Relying on the iconic power of the new ribbon symbol, Google made no reference to September 11 or the attacks; next to the ribbon, they offered a link simply labeled "News and Resources."

It is illustrative to compare how the set of links Google provided in its `/news` folder evolved over the days since the attack. For instance, note

the links offered on September 14 compared to those on October 2:

Google News Resources - September 14, 2001	Google News Resources - October 2, 2001										
 Home - All About Google Current Events	 Home - All About Google News and Resources										
News and information about attacks in US Below are links to news sites and support resources related to the terrorist attacks on the US.	Our thoughts are with all those affected by the recent tragic events. We offers these links as a way of providing resources and up-to-the minute information on developing news stories. We hope you find them useful.										
<table border="0"> <tr> <td> Current News • ABC News • CBS MarketWatch • CNET.com • CNN.com • EuroNews • FOX News • Jane's.com • MSNBC • NPR • New York Post • New York Times • Reuters • Salon • TIME.com • USA Today • Washington Post • Yahoo! News • Yahoo! News Photos </td><td> Support Resources • American Red Cross • America's Blood Centers • American Airlines • United Airlines • Amtrak • Pentagon Updates • Report terrorist activity • Emergency Contact Info (CNN) • Report You're Safe • WTC Survivor Database • Helping.org • FirstGov </td></tr> <tr> <td> • Asahi Shimbun • BBC News • CBC.ca • CNN en Espanol • CNN Japan • Dagbladet.no • Dagens Nyheter • De Telegraaf • De Volkskrant • Diario de Noticias • Die Zeit • elmundo.es • El Pais.es • Guardian.co.uk • Ha' Aretz • International Herald Tribune • La Repubblica.it • Le Monde • Le Figaro • Pravda • Politiken.dk • Saunalahti.fi • Sky News • Spiegel • Svenska Dagbladet </td><td> Relief Funds • Donate via Yahoo! • Donate via Amazon • Donate via PayPal </td></tr> <tr> <td> Contact Information (Airlines & WTC Businesses) • American Airlines: 1-800-245-0999 • United Airlines: 1-800-932-8555 • AON Risk Mgmt Corp: 1-866-256-4154 • Marsh & McLennan: 1-212-345-6000 • Keefe Bruyette and Woods: 1-800-726-0006 • Cantor Fitzgerald: 1-866-326-3188 • Morgan Stanley: 1-888-883-4391 • Fuji Bank: 1-888-537-3845 • Carr Futures: 1-800-755-7620 • Lee Hecht Harrison: 1-201-782-3704 • Thatcher, Proffitt, & Wood: 1-800-698-4567 • Fiduciary Trust Co: 1-800-632-2350 ext. 22578 </td><td> Relief Funds • American Liberty Partnership • Donate via Yahoo! • Donate via Amazon • Donate via PayPal • United Way NYC • Salvation Army • eBay Auction for America </td></tr> </table>	Current News • ABC News • CBS MarketWatch • CNET.com • CNN.com • EuroNews • FOX News • Jane's.com • MSNBC • NPR • New York Post • New York Times • Reuters • Salon • TIME.com • USA Today • Washington Post • Yahoo! News • Yahoo! News Photos	Support Resources • American Red Cross • America's Blood Centers • American Airlines • United Airlines • Amtrak • Pentagon Updates • Report terrorist activity • Emergency Contact Info (CNN) • Report You're Safe • WTC Survivor Database • Helping.org • FirstGov	• Asahi Shimbun • BBC News • CBC.ca • CNN en Espanol • CNN Japan • Dagbladet.no • Dagens Nyheter • De Telegraaf • De Volkskrant • Diario de Noticias • Die Zeit • elmundo.es • El Pais.es • Guardian.co.uk • Ha' Aretz • International Herald Tribune • La Repubblica.it • Le Monde • Le Figaro • Pravda • Politiken.dk • Saunalahti.fi • Sky News • Spiegel • Svenska Dagbladet	Relief Funds • Donate via Yahoo! • Donate via Amazon • Donate via PayPal	Contact Information (Airlines & WTC Businesses) • American Airlines: 1-800-245-0999 • United Airlines: 1-800-932-8555 • AON Risk Mgmt Corp: 1-866-256-4154 • Marsh & McLennan: 1-212-345-6000 • Keefe Bruyette and Woods: 1-800-726-0006 • Cantor Fitzgerald: 1-866-326-3188 • Morgan Stanley: 1-888-883-4391 • Fuji Bank: 1-888-537-3845 • Carr Futures: 1-800-755-7620 • Lee Hecht Harrison: 1-201-782-3704 • Thatcher, Proffitt, & Wood: 1-800-698-4567 • Fiduciary Trust Co: 1-800-632-2350 ext. 22578	Relief Funds • American Liberty Partnership • Donate via Yahoo! • Donate via Amazon • Donate via PayPal • United Way NYC • Salvation Army • eBay Auction for America	<table border="0"> <tr> <td> Current News • ABC News • CBS MarketWatch • Christian Science Monitor • CNET.com • CNN.com • FOX News • IDG News • Jane's.com • Military.com • MSNBC • NPR • New York Post • New York Times • Reuters • Salon • SiliconValley.com • TIME.com • USA Today • Washington Post • Wired News • Yahoo! News • Yahoo! News Photos </td><td> Support Resources • American Red Cross • Helping.org • FirstGov • Greater NY Hospital Association • Coping with disaster • Local Red Cross chapters • People reported missing (CNN) • Photos of missing persons • Report you're safe • Report terrorist activity • Dept. of Justice Fraud Alert • Report civil rights violations • Emergency contact info (CNN) </td></tr> <tr> <td> • Arabic News Online • Asahi Shimbun • BBC News • CBC.ca • CNN en Espanol • CNN Japan • Dagbladet.no • Dagens Nyheter • De Telegraaf • De Volkskrant • Diario de Noticias • Die Zeit • elmundo.es • El Pais.es • Guardian.co.uk • Ha' Aretz • International Herald Tribune • La Repubblica.it • Le Monde • Le Figaro • Pravda </td><td></td></tr> </table>	Current News • ABC News • CBS MarketWatch • Christian Science Monitor • CNET.com • CNN.com • FOX News • IDG News • Jane's.com • Military.com • MSNBC • NPR • New York Post • New York Times • Reuters • Salon • SiliconValley.com • TIME.com • USA Today • Washington Post • Wired News • Yahoo! News • Yahoo! News Photos	Support Resources • American Red Cross • Helping.org • FirstGov • Greater NY Hospital Association • Coping with disaster • Local Red Cross chapters • People reported missing (CNN) • Photos of missing persons • Report you're safe • Report terrorist activity • Dept. of Justice Fraud Alert • Report civil rights violations • Emergency contact info (CNN)	• Arabic News Online • Asahi Shimbun • BBC News • CBC.ca • CNN en Espanol • CNN Japan • Dagbladet.no • Dagens Nyheter • De Telegraaf • De Volkskrant • Diario de Noticias • Die Zeit • elmundo.es • El Pais.es • Guardian.co.uk • Ha' Aretz • International Herald Tribune • La Repubblica.it • Le Monde • Le Figaro • Pravda	
Current News • ABC News • CBS MarketWatch • CNET.com • CNN.com • EuroNews • FOX News • Jane's.com • MSNBC • NPR • New York Post • New York Times • Reuters • Salon • TIME.com • USA Today • Washington Post • Yahoo! News • Yahoo! News Photos	Support Resources • American Red Cross • America's Blood Centers • American Airlines • United Airlines • Amtrak • Pentagon Updates • Report terrorist activity • Emergency Contact Info (CNN) • Report You're Safe • WTC Survivor Database • Helping.org • FirstGov										
• Asahi Shimbun • BBC News • CBC.ca • CNN en Espanol • CNN Japan • Dagbladet.no • Dagens Nyheter • De Telegraaf • De Volkskrant • Diario de Noticias • Die Zeit • elmundo.es • El Pais.es • Guardian.co.uk • Ha' Aretz • International Herald Tribune • La Repubblica.it • Le Monde • Le Figaro • Pravda • Politiken.dk • Saunalahti.fi • Sky News • Spiegel • Svenska Dagbladet	Relief Funds • Donate via Yahoo! • Donate via Amazon • Donate via PayPal										
Contact Information (Airlines & WTC Businesses) • American Airlines: 1-800-245-0999 • United Airlines: 1-800-932-8555 • AON Risk Mgmt Corp: 1-866-256-4154 • Marsh & McLennan: 1-212-345-6000 • Keefe Bruyette and Woods: 1-800-726-0006 • Cantor Fitzgerald: 1-866-326-3188 • Morgan Stanley: 1-888-883-4391 • Fuji Bank: 1-888-537-3845 • Carr Futures: 1-800-755-7620 • Lee Hecht Harrison: 1-201-782-3704 • Thatcher, Proffitt, & Wood: 1-800-698-4567 • Fiduciary Trust Co: 1-800-632-2350 ext. 22578	Relief Funds • American Liberty Partnership • Donate via Yahoo! • Donate via Amazon • Donate via PayPal • United Way NYC • Salvation Army • eBay Auction for America										
Current News • ABC News • CBS MarketWatch • Christian Science Monitor • CNET.com • CNN.com • FOX News • IDG News • Jane's.com • Military.com • MSNBC • NPR • New York Post • New York Times • Reuters • Salon • SiliconValley.com • TIME.com • USA Today • Washington Post • Wired News • Yahoo! News • Yahoo! News Photos	Support Resources • American Red Cross • Helping.org • FirstGov • Greater NY Hospital Association • Coping with disaster • Local Red Cross chapters • People reported missing (CNN) • Photos of missing persons • Report you're safe • Report terrorist activity • Dept. of Justice Fraud Alert • Report civil rights violations • Emergency contact info (CNN)										
• Arabic News Online • Asahi Shimbun • BBC News • CBC.ca • CNN en Espanol • CNN Japan • Dagbladet.no • Dagens Nyheter • De Telegraaf • De Volkskrant • Diario de Noticias • Die Zeit • elmundo.es • El Pais.es • Guardian.co.uk • Ha' Aretz • International Herald Tribune • La Repubblica.it • Le Monde • Le Figaro • Pravda											

New York Hospital Information

- [Greater New York Hospital Association](#)
(List of hospitals in NYC area)
- Mayor's Hotline Center:
(212) 560-2730
(For families seeking to know if their loved ones were seen by a hospital in the NYC area)

- [Politiken.dk](#)
- [rediff.com](#)
- [Saunalahti.fi](#)
- [SME Online](#)
- [Sky News](#)
- [Spiegel](#)
- [Svenska Dagbladet](#)

For instance, note that on September 14, Google offered links to New York hospitals as well as links -- and telephone numbers -- for businesses that had had operations in the World Trade Center. By October 2 such links had been removed, replaced by links to other sites that could provide victim information. The October 2 list improved upon the list of respected international news sources.

The important point to note here is that someone at Google was playing the role of editor, selecting major, authoritative sources, and updating the content to match users' needs as they changed over time. This is an entirely new role for Google.

Google Use in Context

Google was by no means the only online source chosen by the American public and the world at large to learn September 11 news. No doubt millions of users pointed their browsers directly at CNN, or MSNBC, or The New York Times, or any of a growing list of foreign news sources. Indeed, as Google's list of online sources grew, many may have followed links to important news sources outside the United States.

Many were glued to the television for hours of the saturation coverage offered by broadcast and cable news outlets. Some cable channels, such as the Home and Garden channel, replaced programming with a statement of condolences. The Learning Channel substituted a BBC television news feed for its own programming.

In New York City the CBS-TV broadcast affiliate saw a surge in ratings, not only because of the nature of its saturation coverage, but also because its in-town rivals had lost their broadcast towers on the World Trade Center.

Of course, most of us get our information from multiple media. Many of us sought information from a combination of television, on radio, in newspapers, on the Internet, and word of mouth. Within a few days, many Americans suffered from an overload of information -- especially the images of the planes striking the World Trade Center and the Pentagon, repeated over and over again by the television channels.

The tragedy brought about inventive ways of using the Internet and other media to seek and to distribute information:

- Soon after the tragedy, many who had escaped the World Trade Center sought to reach family, friends, or colleagues by telephone. A major Verizon switching center had been destroyed, and long distance lines were saturated. Cell lines also were quickly saturated, so many people desperate to reach loved ones used the text-mode messaging features of their cell phones to communicate with outsiders. For many, this may have been the first practical use of text-mode instant messaging on a telephone.
- New Yorkers and those living in the D.C. area with cable modem, DSL, or direct Internet access in the workplace were able to send e-mail to others inside and outside the New York area. Some who could not get a long distance line were able to dial a local ISP to get the word out.
- Others used the venerable text mode chat tool known as Internet Relay Chat (IRC) to get the word out about their condition in particular and the news in general.
- As the various media sites became overloaded, some informal mirror sites sprang up, with individuals building their own caches of news stories and photographs. Other individuals sent emails with "deep links" to particular pieces of important content on media sites, so that their correspondents would not have to wend their way through slow-painting menu pages.

Within only a couple of days after the attacks, donations had begun surging into various charity sites. Sadly, scam charity sites and e-mails appeared as well. Various bits of disinformation and misinformation began circulating by e-mail, in chat rooms, and on Usenet. A human-edited list of authentic and authoritative sites for information, for donations, and for loved ones to learn about those missing, became critical. This is exactly what Google ultimately provided in its /news folder.



Google home page on Tuesday, September 18 -- one week after the attack.

Click on thumbnail to see full-size image.

Google Responds

I asked Cindy McCaffrey, Vice President of Corporate Communications for Google, to comment on the evolution in the stance Google presented after the attacks. She observed:

We did evolve our site over time as we learned about more information sources and as more sites became accessible. We began placing cached versions of articles on our website early in the day on Sept. 11. We knew that many of the popular news sites were being overwhelmed by people trying to access their sites. As we were able to capture information, we cached the articles and provided a link to the article on our home page.

McCaffrey puts a slightly different spin on Google's initial advice that site visitors turn to traditional broadcast media:

We only recommended that people turn to TV or radio for the most updated information while we had the cached links on our home page (our intent was misinterpreted as a recommendation to not use the Internet, but instead go to the TV or radio; we merely wanted people to understand that if they wanted the most current news, they should check TV/radio). As the news sites were able to handle capacity, we began putting live links to their sites on our home page, and then on our news page linked off the home page.

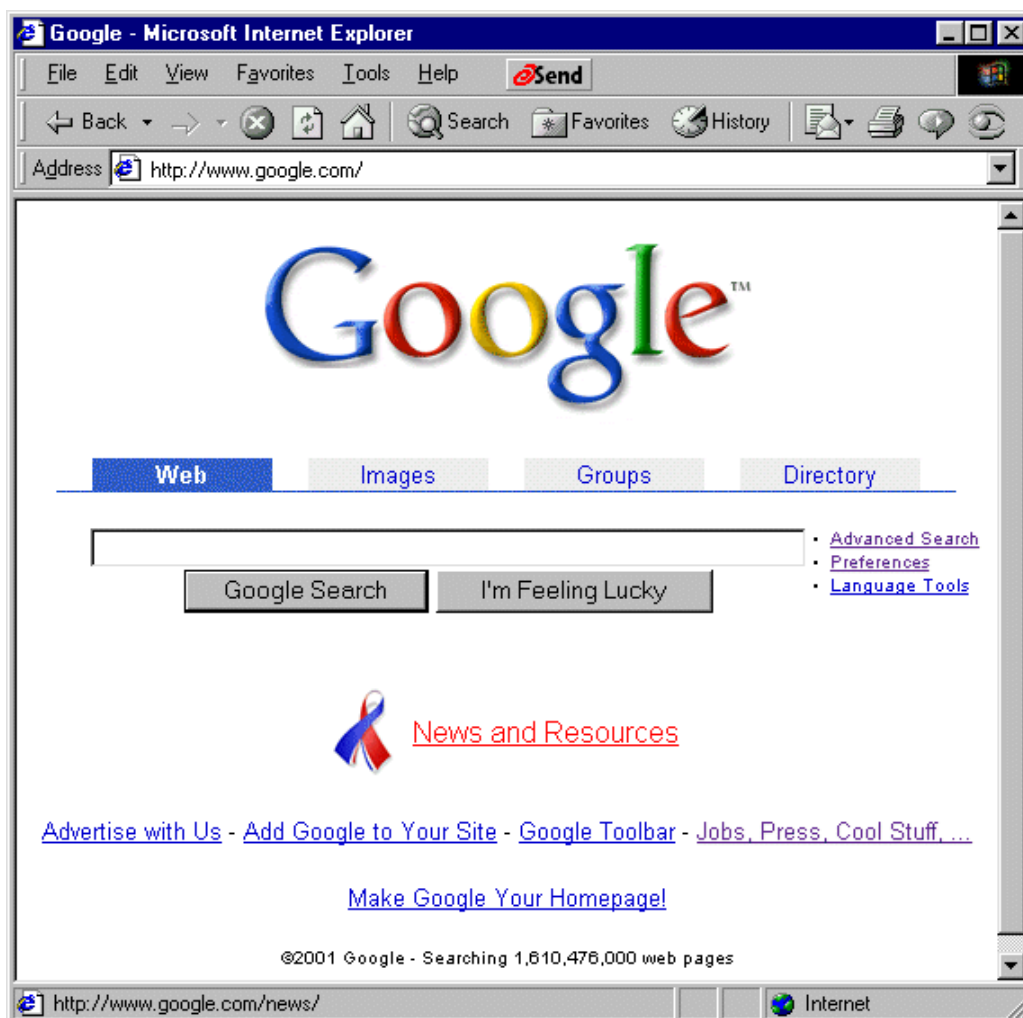
Asked if Google noticed an overall increase in traffic after the disaster, or if there were any negative effects on performance. McCaffrey replied "[There were] no performance issues. Google currently handles about 120 million searches every day. Traffic was not significantly impacted by Sept. 11's events." It is possible that Google has a large body of regular users who simply turned their attention from whatever they normally search for to topics related to September 11. It would be very interesting to know how many of the 120 million queries on September 11 came from occasional users and how many from regular Googlers.

Finally, asked if Google is now committed to populating its `/news` folder, McCaffrey commented: "We don't have any definite plans at this time. We've received positive feedback from people who have indicated that this resource has been helpful to them, but we have not made any decisions on what we'll do with this page longer term."

Where Does Google Go from Here?

Where does Google go from here? Will this disaster set a precedent for the world's leading search engine? Is the "breaking news" section permanent? If so, will Google hire a 24 X 7 staff of editors, enter into content sharing and syndication deals with old and new media partners, and transform themselves into a portal?

Perhaps coincidentally, by early October Google had made a basic change to its front page interface. Tabs appeared, allowing users to switch among basic Web search, image searches, Google Groups (the window to Usenet News built after Google's acquisition of Deja.com), and the Google Directory:



After performing a search, the user can click on one of the other tabs to re-execute the same search in one of the other Google indexes. It's easy to imagine Google building a nearly-real-time index of major news sources, and offering a fifth tab that allows users to search or to browse those sources. The cross-source real time news search concept was pioneered by Totalnews.com and later implemented by Infoseek and Northern Light as specialty searches. Will Google extend its role as the first place millions of Internet users turn to for relevant answers by adding a permanent news section? The answer depends on how well the company understands what people now expect from Google post-September 11.

My guess is that Google will retain a breaking news component on their site -- either on their own devices, or by working with one or more of the major media outlets whose Web searches are "powered by Google." I expect (and hope) that Google will resist the temptation to follow portal failures such as AltaVista, Time's Pathfinder, and Disney's go.com. Google will not clutter their home page -- nor will they lose sight of their vision that search is king. Still, I suspect we have not seen the last "Breaking News" content on Google's site. By meeting user demand for trusted information sources relating to September 11, Google has trained millions of people to expect Google.com to deliver breaking news.

Footnote: A Sadly Ironic Death

September 11 brought tragedy in many ways: the loss of thousands of lives, the loss of an American landmark, the shocking demonstration of how open the United States is to attack. As we have seen, the Internet played an important role in the dissemination of urgently-sought news about the unfolding events. Some aspects of the design of the Internet make it very effective in crisis. An important example is the inherent ability of TCP/IP to re-route and deliver data when important communications links are broken, as in war, or as in the September 11 tragedy. This robustness is not surprising, given that TCP/IP was born out of DoD-funded research.

But in other ways, the Web and the Internet function poorly when millions of people flock to a handful of trusted media sites. Television or radio broadcast distribution is unaffected when CNN's audience goes from 100,000 to 10,000,000. By contrast, nytimes.com can't deliver content when traffic increases a hundredfold: the communications links and the Times' servers are overwhelmed. The Net is still a narrowcast medium, not suited to "mass attention" events.

But the Net is far better at handling millions of users visiting a small number of sites than it was in 1993, thanks in no small part to technologies from a company called Akamai. The problem of mass Web audiences was recognized in the early 1990s, and solutions such as distributed caching and "multicast" streaming were developed. Akamai is a company to

commercialize products aimed at solving the problem. Akamai's products automate what Google did by hand with its cached copies of NY Times articles. Akamai customers include major media companies with Web presences, including CNN. It is sadly ironic that one of the founders of Akamai, Daniel C. Lewin, died in one of the hijacked planes. 

About the Author

Richard Wiggins is a Senior Information Technologist at Michigan State University. He writes frequently about Internet topics for national publications. <http://richardwiggins.com>.

Editorial history

Paper received 18 September 2001; accepted 20 September 2001; revised 3 October 2001.

Contents **Index**

Copyright ©2001, First Monday

The Effects of September 11 on the Leading Search Engine by Richard W. Wiggins
First Monday, volume 7, number 10 (October 2001),
URL: http://firstmonday.org/issues/issue6_10/wiggins/index.html

D. Weaver and B. Bimber (2008).
“Finding news stories: A comparison of
searches using LexisNexis and Google
News,” *Journalism & Mass
Communication Quarterly*. 85(3), 517-
532.

FINDING NEWS STORIES: A COMPARISON OF SEARCHES USING LEXISNEXIS AND GOOGLE NEWS

By David A. Weaver and Bruce Bimber

This study compares Google News to LexisNexis for finding stories in the New York Times, eight large-circulation U.S. newspapers, and all English-language news outlets in each database. Inter-database agreement between Google News and LexisNexis ranged from 29% to 83%, with much of the discrepancy due to wire service exclusions: LexisNexis missed half or more of stories appearing in major papers and in broad searches of English-language news because it is blind to wire stories. Wire-service blind spots in news archives can be a substantial limitation.



Researchers studying media commonly employ digital research archives for data about news. Services such as LexisNexis, NewsBank, ProQuest, and others provide full-content access to many news sources and have long been standard research tools. Recently, free Web-based news portals such as AOL News, Google News, Yahoo! News, and Slashdot have presented potential alternatives to traditional archives, but are designed for public rather than academic research. Research relying on these services is now being reported and shedding new light on some of the limitations of traditional archives.¹

One well-known drawback of traditional news archives is that they typically offer text-only versions of original stories lacking graphics and information about headline size and story placement, which can be important to certain kinds of research. A potentially more serious problem with traditional archives is the fact that content originating with wire services is typically stripped out of newspapers before stories are archived.² This means that news databases do not necessarily constitute archives of the whole content of news appearing in a particular news outlet or in a specific news market. This limitation is of particular concern for research requiring accurate assessments of all news in distribution about a particular subject or in a particular venue, but it is not widely recognized by researchers. It is clear that a good deal of news, especially national news distributed at the local

David A. Weaver is a Ph.D. student in the Department of Political Science at the University of California, Santa Barbara. Bruce Bimber is professor in the Departments of Political Science and Communication at the University of California, Santa Barbara. The authors thank the UCSB Center for Nanotechnology in Society; Lucia Snowhill and Harold Colson for consultation on LexisNexis; as well as Jerry Macala, Erica Lively, Gary Haddow, and Nicole Tyler for research assistance. This material is based upon work supported by the National Science Foundation under grant SES 0531184.

J&MC Quarterly
Vol. 85, No. 3
Autumn 2008
517-532
©2008 AEJMC

level, is comprised of wire stories that are lost in the news archiving process.

To explore this methodological challenge to news research, this study compared a traditional news archive with a Web-based news aggregator: LexisNexis, the most widely used news archive in the social sciences,³ and Google News, which reaches an audience of about nine million people per month and which has been the subject of a good deal of scholarly and professional interest.⁴ The study used the two databases to track a news topic across many months, examining the entire news stream from elite newspapers to local outlets in order to assess the comparability of search results between LexisNexis and Google News, and to use Google News to estimate the number of stories missing from LexisNexis because of omission of wire service content.

As the news business struggles with fragmenting audiences, shrinking advertising revenue, and reductions in news staff, many observers expect wire services, syndicated content, and even blogs and public-contributed material to play an increasingly important role in the news.⁵ This would mean that, absent changes in practice about what is archived, traditional news databases would grow less useful for identifying what news reaches the public.

For the analysis, a topic was chosen that is part of a larger study of news coverage of technological innovation, especially issues of regulation and public safety of new technologies. This study focused specifically on the class of technologies labeled "nanotechnology," which is emerging as a news topic. For example, in 2006, a product recall occurred in Germany that generated news in Europe and the United States. The same year, the Environmental Protection Agency (EPA) announced it would regulate products containing nanotechnology in cases where manufacturers make bacteria-killing claims, and environmentalists and consumer groups petitioned for regulatory action from the Food and Drug Administration (FDA), which declined to act.

These and other developments associated with nanotechnology have two useful characteristics. First, nanotechnology presents a discrete topic with a distinctive and identifiable name, facilitating reliable and valid searches. Second, because the topic is new to journalists and editors, a good deal of variation in attention to it over time and across outlets is expected, and this variation should provide a strong test of how the databases compare at finding news stories.

Methodological Challenges in Use of News Databases

Research on media typically treats the use of archival news databases as unproblematic. This is true of work on agenda setting and other aspects of news and public opinion, of research on independence of news media from government and the relationship of local to national news, and studies of news consonance, among other topics.⁶ However, the methodological literature on use of databases has identified several limitations. Research has shown problems arising from variation between original news content and the content in databases due to truncations of articles, mistakes in capturing headlines, unannounced changes in con-

tracts between news sources and news databases, and idiosyncratic exclusions from databases, such as corrections or retractions from newspapers that are not archived.⁷ Variation of this kind can introduce error into studies of news content, but this error is not large and more importantly does not necessarily bias results systematically.

A more troublesome class of problem involves systematic database exclusions.⁸ Some items are obvious in their absence from text-based databases, such as photographs and other graphics, headlines, and cues about article placement. To find such information, researchers may be able to consult the original news source in print or microfilm, or in many cases the news source's own online archive. A more troublesome exclusion involves wire service stories. Stories originating with services such as the Associated Press are virtually always removed when a newspaper's content is archived in research databases, because licensing agreements customarily permit one-time use only of wire content by a news outlet. The same stipulation typically applies to newspapers' own online archives, which also generally omit wire-originated stories.

This problem was noted by Snider and Janda in an unpublished manuscript in 1998, but has not been assessed systematically.⁹ Research designs that require finding news stories that reached the public—as opposed to studying news production, for example—rarely address the problem. For instance, a recent study of agenda setting in health news used LexisNexis for finding articles in a set of major newspapers, in order to analyze public exposure to positively and negatively valenced news.¹⁰ The technique produced 715 stories. Another search using the authors' own search terms shows that wire services *produced* roughly a hundred articles on the topic during the period under study, some unknown number of which would have appeared in the papers examined, and some of which would have appeared in multiple outlets.¹¹ The Associated Press accounts for over 90% of those stories, meaning that omitted news was dominated by a single organization's perspective.

The fraction of news reaching the public that originates with wire services is not known. Niederdeppe, Frosch, and Hornik estimate that more than 85% of U.S. newspapers subscribe to the AP service, but do not estimate how much news in distribution is actually AP-based.¹² Within a single newsroom, an editor may accept anywhere from a few to half of available stories from the AP.¹³ While large news businesses typically report more stories written by their own staff, local news outlets may rely entirely on wire stories and syndicated material for national news.¹⁴ Some researchers have approached this problem by coding AP content itself as a proxy for national news, but this technique is not common and is limited by absence of information about the distribution of AP stories.¹⁵

The changing structure of the news business may be exacerbating this problem of finding what news the public actually sees. Shrinking audiences for national news broadcasts and falling subscriptions and advertising revenues have led to substantial reductions in news staff at

newspapers across the country; many news businesses appear to be relying more over time on wire stories and syndicated material, as well as Web-based content such as blogs, which are also typically excluded from news archives.¹⁶ One study showed that after purchase by Gannett, more than half of stories in the *Louisville Courier-Journal* originated from wires rather than staff writers.¹⁷

Background on Lexis- Nexis and Google News

LexisNexis and Google News offer a useful comparison for this problem. LexisNexis is the most widely used news archive in the social sciences.¹⁸ For the United States, it offers roughly 300 newspapers and about 500 general print publications, along with transcripts from about three-dozen broadcast outlets. It covers virtually all big-city newspapers in the United States; many papers from mid-size markets, such as the *Charleston Gazette* and *Anchorage Daily News*; and a smaller sample of local news outlets.¹⁹ LexisNexis omits wire service stories in distribution, though it provides searches of wire content as it is produced.²⁰ Methodological analyses of LexisNexis are few in number. The most comprehensive report is that of Deacon, who compared LexisNexis and the Chadwyck Healy newspaper archives for the United Kingdom and reported a number of small to moderate-sized reliability problems due to idiosyncratic differences.²¹

Google News has been a Web-based “portal” to news since 2002, and some of its features are complementary to those of LexisNexis. It is unique among news aggregators in providing a fully automated sweep of global news, conducted roughly every fifteen minutes and without human editorial judgments.²² Google claims to include 4,500 news sources. Its database can be searched in full-text mode, like LexisNexis, and is also represented by a changing home page of some three dozen news topics currently covered by the largest number of news businesses globally.²³ A major news topic, such as an election outcome, might be linked to 5,000 or more stories, while less prominent topics are linked to a hundred or fewer. Though it produces no news of its own, Google News has been described as a paradigmatic example of the changing news business, because it provides “radically expanding access to diverse viewpoints on any issue,” in contrast to journalism’s traditional emphasis on creating filtered, ordered, and prioritized news.²⁴

One of the most useful features of Google News for the present study is the fact that between its inception in 2002 and late summer of 2007, it included wire stories as they appeared in distribution throughout the news system. An AP story, for example, that ran in a hundred newspapers around the country would produce a hundred results, making possible analysis of news distribution. By contrast, that same article would appear once in the LexisNexis database, associated with its producer, the AP. In September of 2007, Google News stopped its practice of displaying “duplicate content,” instead agreeing to license stories directly from four major wire services and thereby reducing a source of complaint about copyright infringement.²⁵ For the five years that Google News provided a window into news-in-distribution, it was possi-

ble to compare results with LexisNexis's database of news-in-production.

Research Questions

The first research question addresses the *New York Times*, which is a common focus of research on news because it is viewed as intrinsically important, or is treated as a proxy for the agenda of the entire news system on the assumption that it serves as a primary "intermedia agenda setter."²⁶ The first step in our comparison examines if searches for *Times* articles with Google News can replicate LexisNexis searches. Then comparison is extended to other large-circulation newspapers in the United States.

RQ1: What is the level of agreement between LexisNexis and Google News when searching for news stories in the *New York Times* and other large-circulation U.S. papers?

The general expectation is that the two databases will produce similar pictures of news coverage, since both databases include the *Times* and roughly the same major papers, and because these papers rely heavily on their own reporting. Any differences will be chiefly the result of idiosyncratic variation in the construction of the two archives.

The second question turns to the contents of the two databases beyond the major newspapers, in order to identify how each portrays news content broadly. Expanding the analysis from major papers to include local news businesses and foreign outlets suggests divergence in results, since LexisNexis's news-production emphasis should result in fewer stories than Google News's news-distribution emphasis.

RQ2: How large is the difference in total number of stories found by LexisNexis and Google News when searching all U.S. and world news?

Even when wire service stories are removed from Google results, some differences should remain in the portraits of the news system provided by LexisNexis and Google News, because they do not search the same body of news sources—Google News reports 4,500 sources on its home page, while LexisNexis has roughly half that number.

In addition to assessing the total volume of news stories that LexisNexis and Google News find, the study compares results when the attention of the media coalesces on a particular issue or development. The assumption is that a baseline level of coverage of the topic will be occasionally interrupted when some key development, such as an action by public officials, leads to a rise in news volume. Without any *a priori* judgment about what should be a newsworthy development, the study compares what LexisNexis and Google News suggest are big news events.

RQ3: What is the level of agreement between LexisNexis and Google News content in depicting major news events?

News Topic: Regulation of Technology. The analysis employs a longitudinal design focused on one topic in order to track coverage in LexisNexis and Google News long enough to observe the rise and fall of various developments in the news, and to see how the two databases handled these. The topic, the safety and regulation of nanotechnologies, offers several useful features for the present inquiry. "Nanotechnology" is a readily searchable topic that produces relatively unambiguous results. As a news topic, nanotechnology is new to most journalists, as it is to the public more generally, yet a number of potentially newsworthy developments occurred during 2006 and 2007, producing a variety of opportunities for reporting.²⁷ In the United States, environmental and other groups were active in advocating for new regulations; several high-level advisory bodies and think tanks issued expert reports about policy; the city of Berkeley, California, adopted an ordinance requiring firms to disclose the presence of nanotechnology within city limits; and the EPA and FDA took preliminary actions on public safety and nanotechnology.

Data and Analysis. To examine how the databases capture the response of news media to such developments, a data set was constructed containing search results for news coverage of nanotechnology in bi-weekly searches from January 29, 2006, to August 15, 2007. This period was chosen in order to begin the dataset prior to the first actions historically by any U.S. regulatory agencies on nanotechnology and also prior to the recall in Germany of a cleaning product called MagicNano. Three months of baseline news prior to these developments were included; data collection continued until about a half-dozen newsworthy events had occurred involving public officials, advocacy groups, or experts. This criterion was met in mid-2007, not long before Google's database change that excluded wire service stories in circulation. The result is an eighteen-month dataset that exceeds the power of traditional news-sample techniques such as the constructed-week sampling approach, and that captures any English-language news found by the two sources.²⁸ The dataset allows plotting of stories over time and calculation of level of agreement between the two.²⁹

The design employed a set of search terms that provide a broad sweep through full text news. After construction of equivalent Boolean constructions in LexisNexis and Google News, searches were performed for stories containing the term "nanotechnology" or several permutations, and at least one term from a list of twenty-one concepts relating to safety, risk, environmental issues, and regulation or public policy, which were developed from the literature on societal implications of nanotechnology and through discussion with topical experts at the National Science Foundation's Center for Nanotechnology in Society at the University of California, Santa Barbara.³⁰ In LexisNexis, the search involved: "General News," which consisted of 52 "Major Papers" and 502 "Magazines & Journals"; "United States News," which consisted of 267 sources from the Midwest, Northeast, Southeast, West, all 50 states, and the District of Columbia; "News Transcripts" which includes material from 35 sources; "World News," which includes 1,289 sources; and "News Wire" sources,

which numbered 357. Accounting for entries appearing in multiple categories, this sample included approximately 2,500 unique news sources. In Google News, parallel searches were conducted of the entire database of 4,500 sources every two to three weeks, in order to stay within the moving window formed by Google's policy of archiving its main news content for only thirty days.³¹ Google's page rank, placement, and other subjective indicators of importance were ignored.

For purposes of examining the major papers including the *New York Times*, the ten largest circulation papers in the United States were identified, using figures from the Audit Bureau of Circulations as of September 30, 2006, which were *USA Today*, *Wall Street Journal*, *New York Times*, *Los Angeles Times*, *Denver Post*, *Chicago Tribune*, *Washington Post*, *New York Daily News*, *New York Post*, and *Houston Chronicle*.³² The *Chicago Tribune* and the *Wall Street Journal* were dropped because LexisNexis does not provide full-text searching of these. The resulting list of eight newspapers constituted the population of news sources for the major-papers comparison.

Each story returned from the searches was coded as relevant or not to the topic, and then, for relevant stories, variables were generated directly from the databases, including *story*, a text string identifying each article by title; *date* as found by the database; *LexisNexis*, a dichotomous variable describing whether the article was found by LexisNexis; *Google News*, which similarly describes whether Google News found the article; *wire*, a dichotomous variable describing whether the article was produced by the Associated Press, United Press International, or another syndicated news service; and *publication*, a text variable giving the name of the news publication.

Reliability. Reliability for coding of raw search results as relevant or irrelevant was assessed, using a simple dichotomous choice about whether each article returned from Google News and LexisNexis searches was on the topic of interest, namely nanotechnology and possible health, environmental, safety, regulatory, or other societal concerns. The reliability procedure for this hand-coded step was as follows. Two coders were used, with about fifteen minutes of initial training. A pilot sample of roughly 50 stories was drawn at random from Google News and coded independently by the coders. The result was perfect agreement, so no further training was employed and the coders proceeded to the main reliability assessment.

The reliability sample has two components. For each database, a week was chosen at random, and all stories from that week were included, up to a target of approximately 150 cases from each database. This sample size permitted 95% confidence or better (two-tailed) in our reliability estimates for each database individually, following Riffe, Lacy, and Fico³³ and assuming actual agreement of 90%. This produced a reliability sample of 304 cases, with 147 from Google News and 157 from LexisNexis. Cohen's kappa for the combined sample was 0.76, 0.79 for the Google News sub-sample, and 0.71 for the LexisNexis sub-sample. These figures are satisfactory for a project of this kind, though not high.

Results

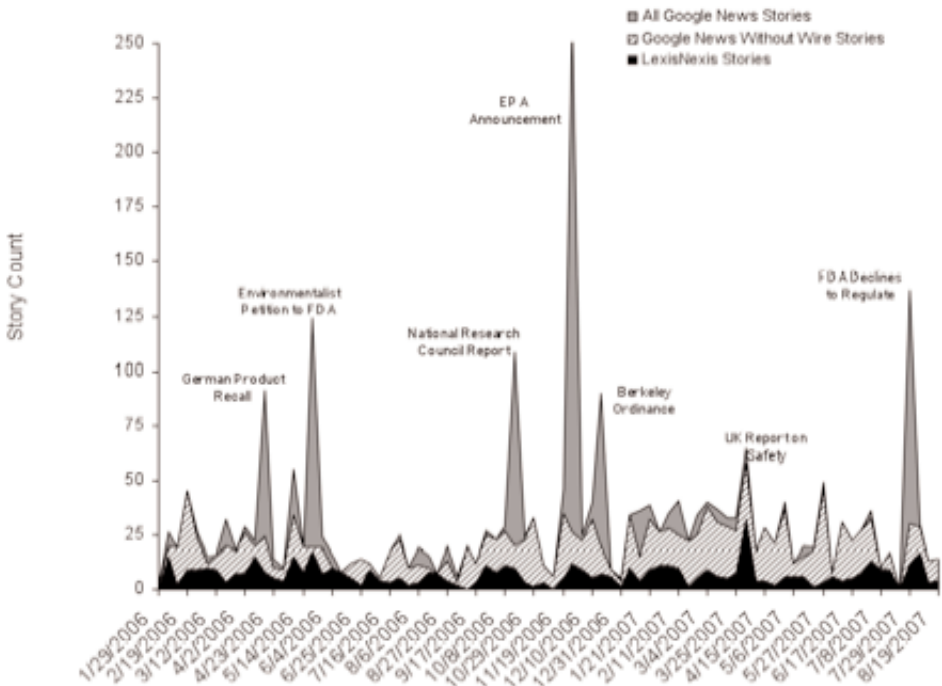
New York Times and Major Papers. The searches returned 2,565 stories from Google News and 558 from LexisNexis used to address **RQ1**. From this set, stories from the *New York Times* were examined. Google News and LexisNexis each returned six stories, five of which they have in common. The common stories covered retail products containing nanotechnology (March 12, 2006); background on societal response to risks from emerging technologies (May 17, 2006); a report by the National Research Council (September 26, 2006); potential regulatory issues associated with food additives (October 10, 2006); and the Berkeley ordinance (January 14, 2007). LexisNexis found an article in the *Times* about an agreement by DuPont and Environmental Defense regarding voluntary safety guidelines (June 21, 2007) that was missed by Google News, while Google found a “Bits” blog entry (August 14, 2007) about potential hazards in sunscreens not included by LexisNexis. Simple agreement was used to compare the two, because the study’s focus is on consistency in the list of stories found by each, rather than in chance-corrected measures that would account for the total number of stories in the databases. For the *New York Times*, agreement was 71% between Google News and LexisNexis.

Next the comparison was expanded to include the rest of the eight large-circulation newspapers. Among these, LexisNexis found 20 articles and Google News 47. These results include 32 articles found by Google News but not LexisNexis, and 5 found by LexisNexis but not Google News. Agreement for these results was 29%. Of the articles found uniquely by Google News, 29 of 32 originated with wire services. Google News thus revealed that LexisNexis’s wire service “blind spot” caused it to miss more than half of the news in the eight major papers, including stories in the *Washington Post*, *Los Angeles Times*, and *Houston Chronicle*. Six of the eight major papers carried the same AP article, a July 2007 piece by Andrew Bridges about the FDA. With these wire stories removed from the analysis, agreement rose to 61%.

All News Sources. **RQ2** involved comparison of the 2,565 stories from all sources returned by Google News to the 558 returned by a search of LexisNexis. Figure 1 plots the number of stories per week found by each. Google News returned an average of 31 stories per week, compared to about 7 for LexisNexis. The difference is consistent with the general expectation, since Google News searches nearly twice as many sources as LexisNexis, in addition to including each occasion that a wire story runs. Google News found 893 instances of news wire stories on the topic, which exceeds the total number of stories found by LexisNexis. Without wire stories, the result was 1,675 Google News stories, for an average of about 20 per week—still more than three times the stories returned by LexisNexis. These are shown in Figure 1 by the line labeled “Google News Without Wire Stories.” There are no obvious patterns in coverage that would explain these differences. For instance, Google News found 10 articles from the *Sydney Morning Herald* missed by LexisNexis, which in turn found 3 in the *Canberra Times* missed by Google News.

FIGURE 1

*Comparing Weekly News Volume Using LexisNexis and Google News,
January 29, 2006, to August 15, 2007,
Stories on Nanotechnology*



News Events. The differences between the two databases persist when one turns to questions of identifying discrete news events, the focus of **RQ3**. In the Google News data, there are a number of spikes in news volume, six of which happen to exceed one standard deviation (35.6) above the weekly mean, and four of which exceeded two standard deviations. Examining the content of stories in these six spikes shows that each reflects coverage of a discrete event, such as the German recall, the petition to the FDA, and an EPA announcement.³⁴ From the perspective of Google News, the biggest news day of the year for the topic was November 23, 2006, when 121 news outlets carried stories about the EPA.

In the LexisNexis data, meaningful news events are less obvious statistically. On nine occasions news volume rises one standard-deviation (4.9) or more above the mean, but inspection of these clusters of stories shows that four do not correspond to any single event or news development, instead reflecting the regular ebb and flow of mis-

cellaneous stories. Four of the six events that trigger one standard-deviation spikes in Google News findings also do so in LexisNexis.³⁵ By this simple approach to identifying spikes in news volume, the databases give quite divergent pictures of what developments were most newsworthy during the study period—that is, of when the attention of media turned more heavily to the topic. Agreement at identifying news events is 36%.³⁶

Discussion and Conclusion

For a topical search exclusively of the *New York Times*, agreement between Google News and LexisNexis is a modest 71%. One of the discrepancies appears idiosyncratic and one is attributable to the fact that Google News includes blog entries. Filtering the blog entry raises agreement to 83%, which approaches an acceptable level and was the highest agreement in the study. From the perspective of inclusiveness of the databases, there seems to be no particular advantage to Google News over LexisNexis for searching the *New York Times*, though some scholars might find Google advantageous because of its link to actual stories in context.

A search of eight major newspapers in the United States yielded poorer agreement (61%) if wire service stories found by Google News are set aside, and only 29% with wire stories in the picture. On this basis, LexisNexis and Google News do not provide comparable portraits of news in major news outlets. Google News revealed that LexisNexis missed more than half of the news stories in major papers. The importance of the wire services to news is not, at least in this case, confined to local coverage of national issues but extends to coverage in major newspapers.

As expected, differences between Google News and LexisNexis for broad searches of U.S. and English-language world news were substantial. Google News found four times as many stories as LexisNexis, and still about three times as many when wire service stories were excluded. For searches of local and foreign news, Google News and LexisNexis should be treated as representing different samples of news sources. So long as Google News keeps its source list proprietary, researchers will not be able to resolve these differences. Google News shows that LexisNexis is blind to a great many news stories because of the wire exclusion, and this problem extends to major news outlets. Caution should be exercised in making inferences from LexisNexis about what news actually reaches audiences. The prominent role of news wires in raising the volume of disseminated stories down the media system is consistent with classic research on news wire services, which finds that wire service news tends to be “uncritically” accepted by local media and by editors operating at national bureaus.³⁷

Which database is better? This study was not designed to render such a judgment for any particular task. Though we regard LexisNexis as the best available news archive, these results are discouraging for researchers who need to identify all stories appearing in a set of news outlets. The problem will grow larger as restructuring of the news business continues and wire services grow more important. Ironically, the

fidelity of news archiving may have diminished since the days when researchers turned to microfilm. Since Google's change in practice in late 2007, both databases now suffer from the wire service blind-spot. It is clear that other techniques should be employed by researchers to account for missing stories when using LexisNexis, such as specifically analyzing AP-produced stories when conducting content analysis of newspaper content, or using selected daily searches of individual news Web sites or print editions in order to estimate the fraction of coverage accounted for by wire content.

Limitations and Areas for Further Research. This study focused on a national news topic low on the political agenda, using Google News and LexisNexis, neither of which is the sole source of data in its class. Our topic has some similarities and a number of differences from other topics of research on news. From the perspective of local news, it is similar to many national issues in that few local outlets cover developments themselves. We would expect the findings about the wire service blind spot to apply in cases of other news topics. From the perspective of national news, the situation may be different. The topic received fairly light and irregular attention from the major papers. The *New York Times* ran only a handful stories, all of which were produced by staff writers, while the rest of the major papers ran a mix of their own reporting and wire service stories. We are hesitant to make inferences from this study to high-salience news topics such as war, political campaigns, or economic news, which might feature stronger gatekeeping, more uniform news agendas, and comparatively less importance of wire service content in coverage by major news outlets.

Several lines of inquiry could improve understanding of these issues. The most important is characterizing the apparently growing reliance of news businesses on wire service and other syndicated content. This could be accomplished through analysis of print news and news Web sites in real time, before archiving processes remove this content.³⁸ Some researchers have employed commercial news clipping services that make daily sweeps through original news, in order to be more inclusive and create datasets with greater fidelity to actual news in distribution, a technique we endorse.³⁹ Comparisons of other news portals, such as Yahoo! News, with other online databases could also improve our understanding of archive biases. In the most general terms, our analysis provides a reminder that any news database provides nothing more than a sample of the actual news to which the public is exposed, and it is important to understand better the nature of sampling effects arising from this fact.

NOTES

1. Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Browstein, "Health Map: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports,"

Journal of the American Medical Informatics Association 15 (March-April 2008): 150-57; Nathaniel D. Poor, "A Cross-National Study of Computer News Sites: Global News, Local Sites," *The Information Society* 23 (2007): 73-83; Tsuyoshi Murata, "Detection of Breaking News from Online Web Search Queries," *New Generation Computing* 26 (November 2007): 63-73; Also see Matt Carlson, "Order versus Access: News Search Engines and the Challenge to Traditional Journalistic Roles," *Media, Culture & Society* 29 (2007): 1014-30.

2. J. H. Snider and Kenneth Janda, "Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis" (paper presented at the annual meeting of the American Political Science Association, Boston, MA, 1998).

3. David Deacon, "Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push-Button' Content Analysis," *European Journal of Communication* 22 (winter 2007): 5-25.

4. Project for Excellence in Journalism, "The State of the News Media 2007: An Annual Report on American Journalism," 2007, <http://www.stateofthenewsmedia.com/2007>; Carlson, "Order versus Access."

5. Project for Excellence in Journalism, "The State of the News Media 2004: An Annual Report on American Journalism," 2004, <http://www.stateofthenewsmedia.com/2004>.

6. Lutz Erbring, Edie N. Goldenberg, and Arthur H. Miller, "Front-Page News and Real-World Cues: A New Look at Agenda-Setting by the Media," *American Journal of Political Science* 24 (February 1980): 16-49; Maxwell E. McCombs and Donald L. Shaw, "The Agenda-Setting Function of the Mass-Media," *Public Opinion Quarterly* 36 (summer 1972): 176-87; Diana C. Mutz and Joe Soss, "Reading Public Opinion: The Influence of News Coverage on Perceptions of Public Sentiment," *Public Opinion Quarterly* 61 (autumn 1997): 431-51; David Weaver, Maxwell E. McCombs, and Donald L. Shaw, "Agenda-Setting Research: Issues, Attributes, and Influences," in *Handbook of Political Communication*, ed. Lynda Lee Kaid (Mahwah, NJ: Lawrence Erlbaum, 2004), 257-82.

7. Snider and Janda, "Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis"; Philip A. Kaufman, Carol Reese Dykers, and Carole Caldwell, "Why Going Online for Content Analysis can Reduce Research Reliability," *Journalism Quarterly* 70 (winter 1993): 824-32; Katheleen A. Hansen, "Online Inaccuracies: The Use and Misuse of Electronic Information Sources" (paper presented at the annual meeting of the AEJMC, Washington, DC, 1995); Daniel Riffe and Alan Freitag, "A Content Analysis of Content Analysis: Twenty-Five Years of *Journalism Quarterly*," *Journalism & Mass Communication Quarterly* 74 (autumn 1997): 515-24.

8. Snider and Janda, "Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis."

9. Snider and Janda, "Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis."

10. Thomas Hugh Feeley and Donald Vincent III, "How Organ Donation is Represented in Newspaper Articles in the United States," *Health Communication* 21 (May 2007): 125-31.

11. We used LexisNexis to search wire services for articles containing [(organ or tissue) and (donation or donor or transplant)] for the period of the study.

12. Jeff Niederdeppe, Dominick L. Frosch, and Robert C. Hornick, "Cancer News Coverage and Information Seeking," *Journal of Health Communication* 13 (March 2008): 181-99.

13. Glen L. Bleske, "Ms. Gates Takes Over: An Updated Version of a 1949 Case Study," *Newspaper Research Journal* 12 (fall 1991): 88-97.

14. Snider and Janda, "Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis." It has also been argued that the Associated Press has become *the* dominant wire service in the news business. See Richard A. Schwarzlose, "The Associated Press and United Press International," in *The Future of News*, ed. Philip S. Cook, Douglas Gomery, and Lawrence W. Lichty (Washington, DC: Woodrow Wilson Center Press, 1992), 145-66.

15. Niederdeppe, Frosch, and Hornik, "Cancer News Coverage and Information Seeking."

16. Project for Excellence in Journalism, "The State of the News Media 2004."

17. David C. Coulson and Anne Hansen, "The *Louisville Courier-Journal's* News Content After Purchase by Gannett," *Journalism & Mass Communication Quarterly* 72 (spring 1995): 205-15.

18. Deacon, "Yesterday's Newspapers and Today's Technology: Digital Newspaper Archives and 'Push-Button' Content Analysis."

19. As context for the size of the LexisNexis database, the 2006 newspaper market in the United States included 1,437 daily papers and more than 6,000 weeklies, and the television market comprised 1,943 licensed broadcast stations. Sources are Project for Excellence in Journalism, "The State of the News Media 2008," 2008, <http://www.stateofthenewsmedia.com/2008>; and Federal Communications Commission, "Broadcast Totals as of December 31, 2006," 2007 (Washington, DC: FCC), http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-269784A1.pdf.

20. Snider and Janda, "Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis."

21. Deacon, "Yesterday's Newspapers and Today's Technology: Digital Newspaper Archives and 'Push-Button' Content Analysis."

22. Carlson, "Order versus Access: News Search Engines and the Challenge to Traditional Journalistic Roles."

23. A 2007 analysis of the twenty-five news sources that appear most commonly among the top Google sources showed that the *New York Times* was the most common source of stories appearing on the Google News home page, followed by the *Washington Post*. Of the ten largest-circulation newspapers in the United States as measured by the Audit Bureau of Circulations, seven appear among the twenty-five most frequent sources of Google News home-page stories, along with several news wires, the major television networks, a number of foreign sources such as the BBC, the *Times Online*, the *Guardian*, and *Xinhua*, and some big-city papers below the ten largest, such as the *Boston Globe* and the

Seattle Post. See Greg Jarboe, "Revealing the Sources of Google News," *Search Engine Land*, May 31, 2007, <http://searchengineland.com/070531-133609.php>. For circulation data, see Audit Bureau of Circulations, "Top 200 Newspapers by Reported Circulation," October 18, 2006, <http://www.accessabc.com/reader/top150.htm>.

24. Carlson, "Order versus Access: News Search Engines and the Challenge to Traditional Journalistic Roles," 1015.

25. Scott Karp, "Google News Hosting Wire Service Stories Diminishes Value of Duplicate Content," *Publishing 2.0*, September 2, 2007, <http://publishing2.com/2007/09/02/google-news-hosting-wire-service-stories-diminishes-value-of-duplicate-content/>, accessed May 20, 2008; Reuters, "Google Shift on Handling of News," *New York Times*, September 1, <http://www.nytimes.com/2007/09/01/technology/01news.html>, accessed May 20, 2008.

26. Frank R. Baumgartner and Bryan D. Jones, *Agendas and Instability in American Politics* (Chicago: Chicago University Press, 1993). See also Yue Tan and David H. Weaver, "Agenda-Setting Effects among the Media, the Public, and Congress, 1946-2004," *Journalism & Mass Communication Quarterly* 89 (winter 2007): 729-44; Maxwell McCombs and Amy Reynolds, "News Influence on Our Pictures in the World," in *Media Effects: Advances in Theory and Research*, 2d ed., ed. Jennings Bryant and Dolf Zillmann (Mahwah, NJ: Lawrence Erlbaum Associates, 2002), 1-18; Maxwell McCombs, Edna Einsiedel, and David Weaver, *Contemporary Public Opinion: Issues and the News* (Hillsdale, NJ: Lawrence Erlbaum, 1991), 47-49. For review of this concept, see Weaver, McCombs, and Shaw, "Agenda-Setting Research: Issues, Attributes, and Influences," 272-73.

27. Dan Kahan, Paul Slovic, Donald Braman, John Gastil, and Geoffrey L. Cohen, "Affect, Values, and Nanotechnology Risk Perceptions: An Experimental Investigation," *Social Science Research Network*, March 7, 2007, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=968652; Chul-Joo Lee and Dietram A. Scheufele, "The Influence of Knowledge and Deference Toward Scientific Authority: A Media Effects Model for Public Attitudes Toward Nanotechnology," *Journalism & Mass Communication Quarterly* 83 (fall 2006): 819-34; Jane Macoubrie, "Informed Public Perceptions of Nanotechnology and Trust in Government" (Washington, DC: Woodrow Wilson International Center for Scholars, September 2005), <http://www.wilsoncenter.org/events/docs/macoubriereport.pdf>.

28. Joe Bob Hester and Elizabeth Dougall, "The Efficiency of Constructed Week Sampling for Content Analysis of Online News," *Journalism & Mass Communication Quarterly* 84 (fall 2007): 811-24.

29. Our data collection procedure missed eleven days of Google News stories during the eighteen-and-a-half-month period: December 29-31, 2006, and July 13-20, 2007. We exclude these from the analysis. LexisNexis returns no articles for these periods from the *New York Times* or the eight major papers, and produces no news events.

30. The search protocol included the following provisions: (a) To identify stories on nanotechnology, we searched in Google News using

"nanoscience"; or "nanoscale"; or "nanotechnology"; or "nano" but not "iPod" or "Apple" or "nanosecond." In LexisNexis, we used "nano!"; or "nano-technology"; and also excluded "Apple" and "iPod" and "nanosecond." All searches examined full text. In the case of Google News, we included stories originally labeled as "omitted" by Google due to potential duplication. (b) To identify stories about politically relevant subjects, we included in the search strings an "and" requirement to at least one occurrence of any from a list of concepts related to politics or political issues.

Because of the absence of wildcard options in Google, we constructed the searches slightly differently. The following is the list of twenty-one pairs of concepts; the first in each pair is the LexisNexis term or terms, and the second, italicized concept is the Google News term or terms. (1) risk!: *risk or risks*; (2) toxic!: *toxic or toxicity*; (3) priva!: *private or privacy*; (4) surveillance: *surveillance*; (5) disast!: *disaster or disasters or disastrous*; (6) goo or self-replicating or swarm or sludge: *goo or self-replicating or swarm or sludge*; (7) ethic! or unethical: *ethics or ethical or unethical*; (8) regulat! or rules: *regulate or regulation or regulations or regulatory or regulators or rules or regulated*; (9) harm!: *harm or harmful*; (10) danger!: *danger or dangers or dangerous*; (11) health or illness or sick: *health, illness, sick*; (12) environment!: *environment, environmental*; (13) standards: *standards*; (14) law! or legal: *law, laws, legal*; (15) policy: *policy*; (16) govern or governance: *governance, govern*; (17) safe!: *safe, safety*; (18) hazard!: *hazard, hazards, hazardous*; (19) controvers!: *controversial, controversy, controversies*; (20) damage: *damage*; (21) concern*: *concern, concerns*.

31. See "Google News: How it Works: Oldest articles in Google News," <http://www.google.com/support/news/bin/answer.py?answer=40215&ctx=sibling&topic=8851>. Beginning in June of 2006, Google added a news archive selectively containing news older than thirty days.

32. Audit Bureau of Circulations, "Top 200 Newspapers by Reported Circulation."

33. Daniel Riffe, Stephen Lacy, and Frederick G. Fico, *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (Mahwah, NJ: Lawrence Erlbaum Associates, 1998).

34. For the Google News events, these are the German recall, labeled in Figure 1 as "Recall;" the petition to the Food and Drug Administration ("FDA Petition"); a report by the U.S. National Research Council ("NRC Report"); release of the FDA advisory report ("FDA Report"); the EPA announcement that it would regulate nanoparticles if germicidal claims are made ("EPA"); and the Berkeley regulation ("Berkeley Ordinance").

35. Exceptions are EPA and NRC.

36. We explored procedures for screening results, such as using a two-standard deviation threshold or removing news spikes not associated with specific events, but were not able to produce higher agreement.

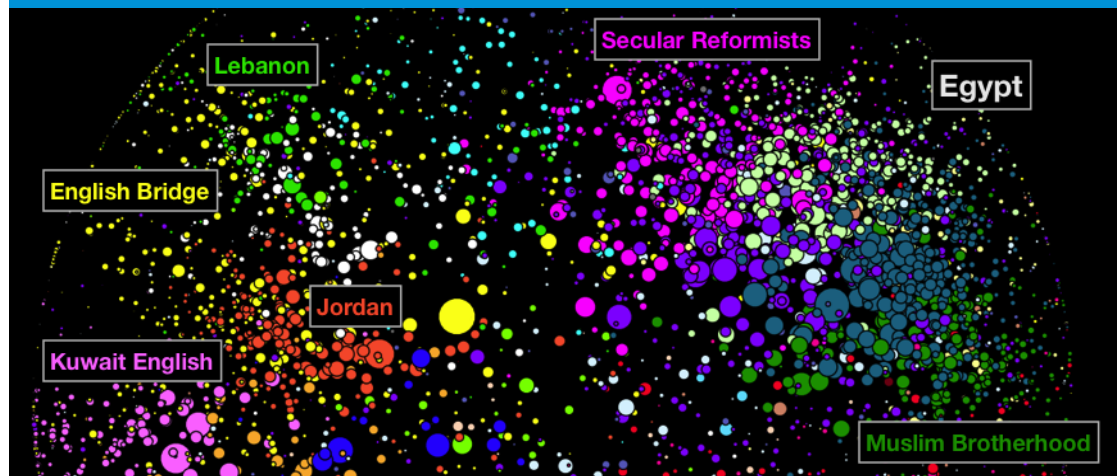
37. D. Charles Whitney and Lee B. Becker, "'Keeping the Gates' for Gatekeepers: The Effects of Wire News," *Journalism Quarterly* 59 (spring

1982): 60-65.

38. For a variation of this approach, see Jeongsub Lim, "Microlevel Agenda-Setting Effects of News Sites" (paper presented at the annual meeting of the International Communication Association, San Francisco, CA, 2007).

39. Katherine Clegg Smith, Melanie Wakefield, Catherine Siebel, Glen Szczypka, Sandy Slater, Yvonne Terry-McElrath, and Sherry Emery, "Coding the News: The Development of a Methodological Framework for Coding and Analyzing Newspaper Coverage of Tobacco Issues," Research Paper Series, no. 21 (Chicago: ImpacTeen, 2002), http://www.impacteen.org/generalarea_PDFs/Newsmethodspaper_smithMAY2002.pdf, accessed May 22, 2008.

B. Etling, J. Kelly, R. Faris and J. Palfrey (2009). "Mapping the Arabic Blogosphere: Politics, Culture and Dissent." Berkman Center Research Publication No. 2009-06, Cambridge, MA: Harvard University.



Mapping the Arabic Blogosphere: Politics, Culture, and Dissent

By Bruce Etling, John Kelly, Robert Faris, and John Palfrey

JUNE 2009

Berkman Center Research Publication No. 2009-06



INTERNET &
DEMOCRACY
project

Digital Methods for Internet Research



Berkman

The Berkman Center for Internet & Society
at Harvard University

ABOUT THE INTERNET & DEMOCRACY PROJECT

This case study is part of a series of studies produced by the Internet & Democracy Project, a research initiative at the Berkman Center for Internet & Society at Harvard University, which investigates the impact of the Internet on civic engagement and democratic processes. More information on the Internet & Democracy Project can be found at:

<http://blogs.law.harvard.edu/idblog/>.

The project's initial case studies investigated three frequently cited examples of the Internet's influence on democracy. The first case looked at the user-generated news site OhmyNews and its impact on the 2002 elections in South Korea. The second documented the role of technology in Ukraine's Orange Revolution. The third analyzed the network composition and content of the Iranian blogosphere. Fall 2008 saw the release of a new series of case studies, which broadened the scope of our research and examined some less well-known parts of the research landscape. In a pair of studies, we reviewed the role of networked technologies in the 2007 civic crises of Burma's Saffron Revolution and Kenya's post-election turmoil. In April 2009, Urs Gasser's three-part case study examined the role of technology in Switzerland's semi-direct democracy. This case expands on our study of foreign blogospheres with an analysis of the Arabic blogosphere.

This paper would not have been possible without the assistance of many individuals. The authors wish to thank our Arabic speaking coders for their tireless efforts reading and interpreting blogs; Anita Patel and Jason Callina for development work on the coding tool; Tim Hwang for research assistance; Lexie Koss for layout and design of the case; Helmi Noman, Noha Atef, and Jillian York for assistance understanding national blogospheres in the region, interpretation of YouTube videos, plus feedback on the draft; and Terry Fisher and Karina Alexanyan for their comments on the draft. Any errors remain our own.

KEY FINDINGS

This study explores the structure and content of the Arabic blogosphere using link analysis, term frequency analysis, and human coding of individual blogs. We identified a base network of approximately 35,000 active Arabic language blogs (about half as many as we found in a previous study of the Persian blogosphere), discovered several thousand Arabic blogs with mixed use of Arabic, English and French, created a network map of the 6,000 most connected blogs, and with a team of Arabic speakers hand coded over 4,000 blogs. The goal for the study was to produce a baseline assessment of the networked public sphere in the Arab Middle East, and its relationship to a range of emergent issues, including politics, media, religion, culture, and international affairs. We found:

- **A Country-Based Network:** We found that the Arabic blogosphere is organized primarily around countries. We found the primary groupings in the Arabic language blogosphere to be:
 - **Egypt** - this is by far the largest cluster and includes several distinct sub-clusters, one of which is characterized by secular reformist bloggers, and another by members of the Muslim Brotherhood, a group that is technically illegal in Egypt but whose online presence appears to be tolerated.
 - **Saudi Arabia** - this comprises the second largest cluster and focuses more on personal diaries and less on politics than other groups.
 - **Kuwait** - this cluster is divided into two sub-clusters based on bloggers' language preferences, splitting those that write primarily in English from those that use Arabic. Both groups focus heavily on domestic news and politics, though the Anglophone bloggers are more likely to advocate reform and discuss economic and women's rights issues.
 - **Levantine/English Bridge** - this group of bloggers is located mainly in the countries of the eastern Mediterranean sometimes referred to collectively as the Levant, including Lebanon, Palestine, Jordan and Syria, as well as most bloggers from Iraq. Bloggers in this group frequently use English in addition to, or instead of, Arabic. They are joined in this section of the network, which connects to the US and international blogosphere, by 'bridge bloggers' from other Arab countries, who write mainly in English.
 - **Syria** - this cluster features frequent, though often mild, criticism of domestic leaders and both includes Arabic language bloggers with closer links to those in Saudi Arabia, and English language bloggers closer to those in the Levantine/English Bridge region.
 - **Maghreb/French Bridge** - this group is comprised of a cross-national set of bloggers located mainly in Morocco, Tunisia, and Algeria. Many of these bloggers write in a mixture of Arabic and French. Other bloggers from the Maghreb eschew French and are found among the religion-focused bloggers.
 - **Islam-focus** – this cluster is a loosely connected set of bloggers from various Arab countries who are focused mainly on Islam, mixing personal, theological, and political topics.

- **Who are Arabic Bloggers?** Demographic coding indicates that Arabic bloggers are predominately young and male. The highest proportion of female bloggers is found in the Egyptian youth sub-cluster, while the Maghreb/French Bridge and Syrian clusters have the highest concentration of males.
- **Personal Life and Local Issues are Most Important:** Overall, the writing of most bloggers is centered on personal, diary-style observations. Those that write about politics tend to focus on issues within their own country and are more often than not critical of domestic political leaders. Foreign political leaders are discussed less often and most commonly in terms more negative than positive. Domestic news is more popular than international news among general politics and public life topics, especially within large national clusters writing entirely in Arabic. The one political issue that commands the most attention of bloggers across the Arab world is Palestine, and in particular the situation in Gaza (Israel's December 2008/January 2009 military action occurred during the study).
- **Discussion of the United States:** The United States is not a dominant political topic in Arabic blogs; neither are the wars in Iraq or Afghanistan. Most discussion of the United States is in the English Bridge and Syrian clusters, and when the United States is discussed, it is usually in critical terms.
- **Muslim Brotherhood:** Although the Muslim Brotherhood is technically illegal in Egypt, it has a very active presence in the blogosphere. Like other Egyptian political bloggers, 'Brother bloggers' talk about human rights and defend those who have been arrested by the government. They are also engaged in public debate about the organization's future and priorities. This debate is mainly between the older establishment leaders and younger reform-minded members, but there is also a wing of younger conservatives pushing back against the latter. Though a minority, a number of Muslim Brotherhood bloggers are women.
- **Terrorism:** Terrorism is a bigger issue among Levantine/English Bridge and Syrian bloggers than others, where it is not a major issue. Across the map however, when discussing terrorism, Arab bloggers are overwhelmingly critical of violent extremists. We consider this a positive finding, although qualified because the issue of attitudes toward terrorism hinge on the term's interpretation across the Arab world. Whatever its presence in other, less 'public' online venues, overt support for violent global confrontation with the West appears to be exceedingly rare in blogs. However, it is not unusual to find blogs that criticize terrorists on the one hand, and praise Hamas or Hezbollah for violent 'resistance' to Israel on the other. This complex issue merits additional research.
- **Iraqi bloggers:** While much has been made of Iraqi bloggers during ongoing debates about the Iraq war, this group does not figure prominently in the Arabic blogosphere. Rather, they are deeply integrated into the English Bridge group. This may be because many Iraqi bloggers write in English and have many inbound links from US think tanks, journalists, and partisan political bloggers (*Iraq the Model* on the

right, *Riverbend* on the left, for example), rather than mainly writing for a domestic public.¹

- **Religion:** Religion is a very popular topic in the blogosphere, and appears to be discussed more in terms of personal religious thoughts and experiences than in its political or theological aspects. Criticism of other faiths is minimal, though can be strident. The exception is a cluster of bloggers within the Islam-focus area, from various countries, who write about Islam from a conservative perspective and frequently criticize other faiths.
- **Human Rights and Culture:** Human rights (civil and political rights) are also a popular topic of conversation across the Arabic blogosphere—much more uniformly common than discussion of Western culture and values, which is concentrated mainly in the Levantine/English Bridge cluster. Among cultural topics, poetry, literature, and art are more popular across the board than pop culture (music, TV, and movies).
- **Arabic Media Ecosystem:** Bloggers link to Web 2.0 sites such as YouTube and Wikipedia (both English and Arabic versions) more than other sources of information and news available on the Internet. Al Jazeera is the top mainstream media source, followed by the BBC and Al Arabiya, while US-government funded media outlets like Radio Sawa and Al Hurra are linked to relatively infrequently. National media outlets are cited mainly by their respective national clusters.
- **YouTube:** Arabic bloggers tend to prefer politically-oriented YouTube videos to cultural ones. Videos related to the conflict in Gaza and the throwing of shoes at George Bush in Iraq are popular across the entire blogosphere, while clips related to domestic political issues are linked to more heavily by the various national clusters.
- **Anonymity:** In general, Arabic bloggers are more likely than not to use their name when blogging, as opposed to writing anonymously or using an obvious pseudonym. The exceptions are Syrian, Kuwaiti, and Maghreb/French Bridge bloggers, as well as a small cluster of Egyptian Baha'i bloggers, who write about their persecution. Generally, women are more likely to blog anonymously than men.

¹ Marc Lynch, "Blogging the New Arab Public," Arab Media & Society, Issue 1, Spring 2007, <http://www.arabmediasociety.com/?article=10> (accessed May 19, 2009).

TABLE OF CONTENTS

Key Findings	3
Introduction	7
Structure of the Network and Methods Overview	12
Results: Network Map Overview	14
Cluster Descriptions	15
Egypt	15
Levantine/English Bridge	19
Maghreb/French Bridge	23
Saudi Arabia	24
Syria	25
Kuwait	26
Islam Focus	28
Baha'i	30
Results: Human Coding	31
Topical and Issue-Based Coding Results	32
Demographic Results	36
Gender and Blogs	37
The Arabic Language Media Ecosystem	38
A Closer Look at YouTube	40
Discussion and Conclusion	46
Appendix A: Notes on Methodology	51
Appendix B: Code Sheets	53

INTRODUCTION

On January 17, 2009, a popular Saudi TV sports program covered the day's disappointing loss by the national soccer team to Oman in the Gulf Cup. The show's host and his guests, including a professional soccer player and a former coach, are critiquing the team and its management when a call comes in from Saudi Prince Sultan bin Fahd, a key patron of the team. He is not happy with their analysis. On air, the prince dresses them down in turn, and goes so far as to tell one of them he is poorly raised, a serious insult in Saudi culture. The prince's tone is disrespectful, and his words are not those of a leader to citizens, but of a ruler to his subjects. A clip of the tirade quickly appears on YouTube, and blogs and online forums post the link, spawning long chains of comments. These are overwhelmingly critical of the prince, who was seen as speaking to the commentators as though they were his slaves.

On August 28, 2008, a clip entitled 'The Pasha's Daughter is Terrifying People on the Street' was uploaded to YouTube. Shot with a mobile phone camera, it shows a young woman threatening a man in the streets of Giza, in the aftermath of a fender-bender in which she hit his car. She brandishes an electric cattle prod and threatens him with the power of her father, a high-ranking security official, whom she calls on her mobile phone. A bystander who captured the event sent the video to a prominent blogger, who posted it on YouTube and tracked down the woman's license plate number. The clip was cited by a number of other Egyptian bloggers, and spawned threads of comments critical of the woman's abuse of her father's power. Other online comments were more critical of the language she used, which was deemed impious. Someone, presumably not Egyptian, re-uploaded the clip under the title 'An Egyptian Woman Insulting God in the Street.' An Egyptian newspaper picked up the story and located the victim, who said that in the aftermath, the woman's mother had come forward to apologize to him, which he accepted.

These two stories (and there are many more) illustrate the collision of old realities and new technologies taking place in the Arab world, and a surprising number of elements intertwine in them: abuse of power, legitimacy of authority, the power of television, the ubiquity of video cameras, feedback between blogs and the press, traditional vs. modern sensibilities, freedom of expression, the power of online voices, and the scope of political arenas—local, national, pan-Arab, pan-Muslim, global. At stake in this collision are both the symbolic construction and the hard power of 'The Public' across the region. Notable is the seamless combination of modes of communication into a single system: face-to-face interaction (including cattle prods), mobile phones, television, newspapers, and multiple genres of Internet sites (blogs, forums, chat rooms, video sharing, photo sharing, etc.). Increasingly, these comprise an emerging *networked public sphere*, in which the power of elites to control the public agenda and bracket the range of allowable opinions is seriously challenged.

Around the world, in open and repressive nations alike, Internet-based communications provide new channels for citizen voices, minority viewpoints, and political mobilization, and challenge the traditional regimes of public mass communication. Opinions are mixed about the promise and dangers that may result. Yochai Benkler presents a compelling view of the networked public sphere as a boon for individual autonomy and freedom, breaking elite strangleholds on democratic discourse and drawing diverse interests and talents into a

common arena.² Cass Sunstein warns of the radicalizing tendencies of micro-publics dividing themselves into groups of the like-minded.³ Public discussion of the Internet in the Arab world often reflects these competing views, with fear of jihadist radicalization on the one hand and hopes for a democratic transformation of the Middle East on the other. Others note the significant advantages that dictatorial regimes retain in controlling the Internet and limiting online speech.⁴ This last point is crucial, since the degree of freedom allowed for online speech may determine whether Benkler's dream or Sunstein's nightmare better describes the Arabic Internet's future.

We looked at a new, and still small, component of the Arabic Internet, but one that in many places has emerged as arguably the most consequential in remaking public sphere communications: the 'blogosphere.' In the evolution of Internet communication genres, older forms like e-mail, listservs, bulletin boards, chat rooms, online forums, and threaded newsgroups have been joined and often supplanted in popularity by newer ones like Web logs (blogs), social network services and, lately, Twitter. In the United States and many other places, these new genres have developed in two principal directions: a) Toward platforms for public speech that foster an interaction among viewpoints and interface with the traditional mass-mediated public sphere, and; b) Toward better-networked private sphere platforms where individuals cultivate their social networks, but also where the like-minded can collect and mobilize mainly outside the view of the uninvited. Private sphere networks such as Facebook can have a powerful public impact, when used to mobilize campaign volunteers in the United States for instance, or a labor strike in Egypt, but the network of persistent, hyperlinked discourse formed mainly by blogs has become an important component of the democratic public sphere in countries like the United States and South Korea, and even in repressive quasi-democracies like Iran.

Older genres of online discourse, particularly forums and password-protected bulletin boards, are extremely popular in the Arab world, and have a history as major venues for politically sensitive topics of all sorts, from women's rights to violent extremism. As more clearly *public* genres of online speech emerge elsewhere, these 'shadow publics' may continue to thrive in part because of government repression (or social sanction) that threaten those who would speak their minds openly. These 'shadow publics' are not in the scope of this study. We fully expect the tone and breadth of discussions in these alternatives forums, which are recognized as active and important venues for Internet activity in the Arabic speaking world, to be substantially different from the blogosphere. We do not intend to either underemphasize the importance of these forums or to imply that observations from the blogosphere are of equal applicability to these spaces.

² Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedoms* (New Haven: Yale University Press, 2006) 129-272.

³ Cass Sunstein, "The Daily We," *Boston Review*, Summer 2001, available at <http://bostonreview.net/BR26.3/sunstein.html> (accessed November 26, 2007); See also Cass Sunstein, *Republic.com* (Princeton: Princeton University Press, 2001).

⁴ Shanti Kalathil and Taylor Boas, *Open Networks, Closed Regimes: The Impact of the Internet on Authoritarian Rule* (Washington, DC: Carnegie Endowment For International Peace, 2003); see also Kristin Lord, *The Perils and Promise of Global Transparency: Why the Information Revolution May Not Lead to Security, Democracy, or Peace* (Albany: SUNY Press, 2006).

Social theorists, from Mill and Dewey to Benkler and Sunstein, stress the importance of open dialog between differing perspectives as a key practice underlying democracy, without which democratic institutions will be ineffective and perhaps short-lived. A key question then in the Middle East is whether a space for open democratic discourse can open up between repressive, elite-controlled public spheres on the one hand, and the hidden discourse of myriad private spheres on the other. The goal of the current study is to examine the Arab publics that are coalescing online in Web logs in order to establish benchmarks for the current state of the blogosphere and to identify issues to watch going forward.

We found a complex network that includes bloggers from at least 18 Arabic speaking countries, as well as Arab expats living in Europe, the United States, South America, and elsewhere, based on large-scale social network mapping and computer text analysis of over 6,000 of the most-linked to Arab Web logs, combined with human coding of over 4,000 blogs. Within the overall network, coherent clusters of blogs emerge, based on similarities in linking choices by bloggers. In the Arabic blogosphere the primary centers of gravity are national, with Egypt, Saudi Arabia, and Kuwait having the largest clusters. Another part of the network draws together bloggers from many nations, especially from the Levant region, in an extensive group that mixes English and Arabic language content in various proportions, and bridges between the Arabic and international blogospheres. A similar but considerably smaller region bridges Arabic to French, mainly through Maghrebi (North African) bloggers. Finally, we see a loosely connected region of bloggers from various countries that focus their attention on Islam, and a small group of Egyptian Baha'is.

Across this diversity of nationalities, we found a wide range of perspectives and topics of concern. While there are some notable concentrations of bloggers around particular points of view, such as Egyptian secular activists, Egyptian Muslim Brotherhood members, and Kuwaiti reformists, it is hard to imagine any issue or perspective in the Arab world for which we did not find a least an advocate or two, including a sympathetic view of Israel's struggle with Hamas (though this was indeed rare). We found Islamists, secularists, and avowed atheists. Among the Islamists there were Salafist Sunnis, Twelver Shi'as, and moderate modernizers. Among secularists there were Western-leaning democrats, anti-Western Socialists and Communists, and a healthy dose of Feminists. Some topics of concern, like women's rights and the Israeli/Palestinian struggle, are discussed by large numbers of bloggers across the map. Other topics, like disputes between Sunni and Shi'a, corrupt royals, or parliamentary politics, were more active in national contexts. Still other topics, like gay rights and atheism, appear to have their niches. Perhaps the most interesting thing was the way in which, as we see in the US blogosphere, the identities and attitudes of real people regularly fail to conform to stereotypical expectations. For a start, many Muslim Brotherhood bloggers are women. And outside the Brotherhood, there are women who support hijab and write of their religious devotion, but also defend their right to pursue attractive men rather than wait passively to be chosen. An educated young Lebanese Christian woman defends her nationalist credentials and also her support of Hezbollah. A Saudi blogger criticizes permissive television programming during Ramadan, and yet posts a provocative picture of a woman. A religious blogger adapts Western self-help ideas to promote better personal discipline in following Islam. Finally, while the range of political perspectives is remarkable, it is important to note that the majority of blogs in most parts of the network are not mainly political, but personal journals which touch on politics to varying

degrees. Mostly, they recount personal events, thoughts, and feelings, frequently touching on themes of love, marriage, jobs, and faith, especially the challenge of finding or preserving any of the first three. As we found in a prior study of the Iranian blogosphere, poetry is very common.

Our findings lead us to push back against two common threads of speculation about the role of the Internet in the Arab world. The first is a view of the Internet as primarily a vehicle for radicalization. We found very little support for terrorism or violent jihad in the Arabic blogosphere and quite a lot of criticism. As discussed above, and again later, a more fully articulated view of Internet-based publics is required, which distinguishes between the role of 'public publics' that participate openly in the public sphere and 'private publics' of the like-minded, but it appears that so far support for violent extremism does not thrive in the former. The second thread is the hope that the Internet will empower political movements that can move Arab societies toward democratic values and governance. This study supports some aspects of the view that the Internet can empower political movements in the region, since it provides an infrastructure for expressing minority points of view, breaking gatekeeper monopolies on public voice, lowering barriers to political mobilization (even if symbolic), and building capacity for bottom-up contributions to the public agenda. But we caution against the idea that these functionally 'democratic' characteristics necessarily travel in the company of classically liberal, historically Western, values regarding individual liberties, freedom of belief and expression, and, ultimately, secular foundations of political rights. The Internet lays a good foundation for a battle of ideas, but it does not necessarily favor a winner.

The results of our study lead us to appreciate the work of analysts who have looked closely at the actual role of bloggers in the Arab world, and have staked out important priorities for research. Key issues include:

- The extent to which Arab online publics will organize into meaningful pan-Arab discourse communities versus reinforcing national divisions [Lynch];⁵
- The specific role of online media in political mobilization, public discourse, and political outcomes [Lynch, Isherwood];⁶
- The implications of the Internet for Islamist movements, including the internal politics of Islamist groups, their capacity to mobilize, and their place in the wider public sphere [Ajemian];⁷

⁵ Marc Lynch, "Blogging the New Arab Public."

⁶ Ibid, see also Tom Isherwood, "A new direction or more of the same? Political blogging in Egypt," *Arab Media & Society*, Issue 6, (Fall 2008), http://www.arabmediasociety.com/articles/downloads/20080929144105_AMS6_Tom_Isherwood.pdf (accessed May 19, 2009).

⁷ Pete Ajemian, "The Islamist opposition online in Egypt and Jordan," *Arab Media & Society*, Issue 4 (Winter 2008), http://www.arabmediasociety.com/topics/index.php?t_article=183 (accessed May 19, 2009) as well as Marc Lynch, "Young Brothers in Cyberspace," *Middle East Report* #245, Winter 2007, available at <http://www.merip.org/mer/mer245/lynch.html> (accessed May 20, 2009).

- The tension between bloggers serving domestic audiences and ‘bridging’ internationally [Zuckerman].⁸

We hope that this study will contribute to addressing these and other issues. In the sections that follow, we present our methods for this study, review several sets of findings, and discuss implications and conclusions with an eye toward the importance of Web logs versus other genres of online discourse.

⁸ Ethan Zuckerman, “Meet the Bridgebloggers,” *Public Choice*, January 2008, Vol. 134, Issue 1/2.

STRUCTURE OF THE NETWORK AND METHODS OVERVIEW

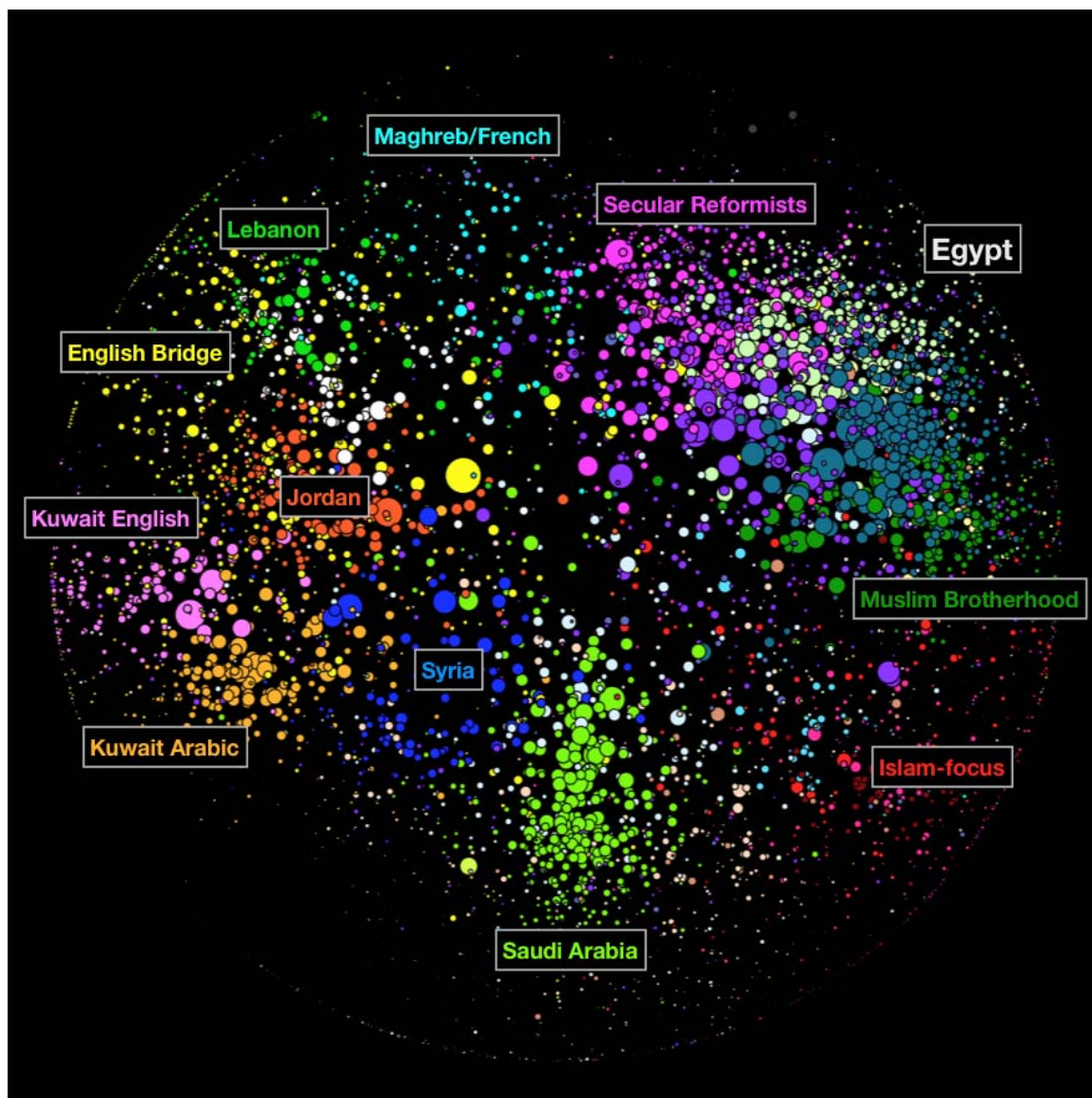


Fig. 1: Map of the Arabic Blogosphere

Adopting similar methods to our study of the Iranian blogosphere, we leveraged content analysis of blogs against a large-scale social network analysis of the Arabic language blogosphere.⁹ Figure 1 is a network map of the Arabic blogosphere, in which each dot represents a blog. The size of the dot represents the number of other blogs that link to it, a measure of its popularity. The position of each dot is a function of its links with its neighbors. The diagram is drawn with a Fruchterman-Rheingold ‘physics model’ algorithm. Imagine that there is a general force trying to push all blogs away from each other, like a

⁹ John Kelly and Bruce Etling, “Mapping Iran’s Online Public: Politics and Culture in the Persian Blogosphere,” Berkman Center Publication No. 2008-01, April 6, 2008, available at http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public (accessed December 6, 2008).

wind blowing them off the map, and that a spring (or force of gravity) pulls together any two blogs that are linked online. Blogs will thus be drawn together by their direct links, but more importantly by the links among their shared neighbors. Thus large groups of blogs cluster up into densely interlinked network neighborhoods. Blogs that share a lot of common neighbors will be close together in the map, even when they do not link directly to one another.

The color of each dot on the map is determined through a different process. A large proportion of the links from blogs are not to other blogs, but to online news sources, organizations, businesses, and various other Internet resources. In some areas of the blogosphere, particularly clusters attentive to news and public affairs, the majority of links are to things other than blogs. On this map, color indicates the assignment of a blog to a particular *attentive cluster*, which is a group of blogs that link to similar online resources. Attentive cluster analysis looks at the full range of online resources linked to by blogs, and groups each blog with others that share a similar linking profile. In blogs, linking to another online resource represents a hard measure of the blogger's attention and serves as the basis for creating attentive clusters, which is accomplished using statistical correlations in linking patterns among bloggers. We see clear patterns in what bloggers are paying attention to across the network. Network neighborhoods and attentive clusters are highly correlated and therefore generally overlap a great deal in network maps. However, it is important to understand that they represent two different things. The first is the direct linking among bloggers, i.e. who is close to whom in the online newsroom, the second is where they are linking to in the Internet at large, i.e. who is looking at the same things, statistically speaking.

The analytic approach of this study is based on the principle that macro structure arises from the tendency of individuals to link more frequently to things they are interested in. This phenomenon is an extension into the blogging world of social behaviors that are well understood in other social scientific contexts. Sociology has an extensive literature on *homophily*, the tendency of social actors to form ties with similar others.¹⁰ Communications research has identified complex processes of *selective exposure*, by which people choose what media to experience, interpret what is experienced, and remember or forget the experience according to their prior interests and beliefs.¹¹ Online behavior is conditioned by the user's preferences in conscious and unconscious ways. In the blogosphere, these preferences express themselves as choices about what to read, write about, link to, and comment upon. As Yochai Benkler notes, linking to those you agree or disagree with is at the very core of blogging and is inherently easy in blog platforms. Linking has contributed to a 'show me' culture on the Internet where readers see for themselves the sources of information a blogger has referenced and make their own judgments about the validity of the source. This

¹⁰ Miller McPherson, Lynn Smith-Lovin, and James M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology* 27 (2001): 415-445.

¹¹ David O. Sears and Jonathan L. Freedman, "Selective exposure to information," *Public Opinion Quarterly* 31 (1967): 194-213; See also D. Frey, "Recent research on selective exposure to information," in *Advances in Experimental Social Psychology*, Vol. 19, ed. L. Berkowitz, (New York: Academic Press, 1986); as well as S.H. Chaffee and Y. Miyo, "Selective exposure and the reinforcement hypothesis: An intergenerational panel study of the 1980 presidential campaign," *Communication Research* 10 (1983): 3-36.

is partly what makes the Internet fundamentally different from the old mass media model, where experts or gatekeepers summarized events, government reports, or research for a mass audience.¹² The social norm among bloggers is to link frequently in one's blog roll and individual posts to influential sources and friends, often with the belief that frequent linking to other bloggers will increase one's visibility in the network. The result of this online linking behavior is a *discourse network*, tractable to empirical research as a massive corpus of text and hyperlinks created by millions of people and stored on thousands of the world's Internet servers. By mapping this network and using the map to view these blogger's communicative activity, we achieve a graphical depiction of the blogosphere that might be described as an fMRI of the social mind.

It is important to note that bloggers that are not linked to frequently by other bloggers are not included in our analysis. In places where state mandated filtering is in effect, linking to certain blogs and sources may increase the likelihood of a blog being blocked, which introduces friction into the link economy. Even in states such as Egypt that seek to inhibit online speech without resorting to filtering, linking to certain blogs and online sources might lead to unwanted attention from government authorities. This may have an impact on the shape of the networked public sphere as represented in this study.

Once the map is created and the network is divided into various attentive clusters, we use text-mining techniques to gain insight into what topics and issues interest each of the clusters. Across the clusters, we analyzed the relative frequencies of bloggers' use of words extracted from all text content, key phrases assembled by the researchers, and known concepts and entities extracted from Arabic language Wikipedia. By looking at which of these were used much more frequently by particular clusters, we are able to make inferences about major topics of concern.

A team of Arabic speaking researchers also conducted exploratory human coding of thousands of blogs to supplement automated techniques and to provide more qualitative assessment of Arabic blogs. Coders were asked to try to ascertain basic demographic data about bloggers (age, location, etc.) as well as blogger opinions on political, cultural, and religious topics. Finally, bloggers provided a short text description of what was unique or stood out about each blogger.¹³

RESULTS: NETWORK MAP OVERVIEW

Several interesting structural features are apparent in the network. First, the Arabic blogosphere manifests several large national clusters. Egypt is by far the largest structural group, and contains several attentive clusters within it. Saudi Arabia and Kuwait also have large, distinct national blogospheres. Beyond these, we continue to find attentive clusters for a number of countries, but many of these, principally the Levant countries, meld into a larger grouping in which we find mixed Arabic and English blogging. This region forms a 'bridge' to the international English language blogosphere, including American blogs. Some national bloggers (like Iraqis) are merged almost completely into the bridge, while others (like Lebanese) have more national coherence. The Syrians are unusual in that some of them,

¹² Yochai Benkler, *The Wealth of Networks*, 216-219.

¹³ See Appendix A for additional information on the methodology.

often blogging in English, merge into the bridge, while others, blogging in Arabic, connect up with the Saudis. Two additional features complete the picture. First, a number of Maghrebi (North African) bloggers connect in a Francophone mirror of the English bridge. Bloggers in the French bridge mix Arabic and French, and link frequently to French Web sites. Lastly, there is a loosely connected, multinational group of bloggers focused on religion, predominantly Islam, though the religion zone also contains a small cluster of Baha'i bloggers as well. The religion zone contains several attentive clusters, mainly reflecting choice of blog hosting service. One religion-focused cluster however contains mainly Islam-oriented Maghrebis, some of whom specifically oppose the use of French by other Magrebi bloggers. Another contains conservative Islamic bloggers focused much more on theology, interpretation of the Qu'ran, and criticism of other faiths. At the end of the cluster descriptions we have also included comments on various blogs made by the coders to provide a more qualitative interpretation of bloggers in each cluster.

CLUSTER DESCRIPTIONS

In this section, we provide the results of various analyses used to interpret the Arabic blogosphere map and the various attentive clusters found in it.

Egypt

Egyptian bloggers comprise the largest structural cluster in the Arabic blogosphere, undoubtedly in part because Egypt has a large online population, and the largest population of any Arab country. Marc Lynch argues that Egypt has some of the most active political bloggers in the Arab public sphere, that is, bloggers who are connected to political movements and are actively engaged in politics, instead of merely discussing politics.¹⁴ Our findings seem to support this argument, especially in the Secular Reformist and Muslim Brotherhood clusters. Egyptian bloggers frequently use pictorial badges on their sites to show support for various campaigns, such as for freeing bloggers, calling for reform, or promoting social issues such as combating HIV/AIDS stigmas. And some of these campaigns, like 'Stop the Imprecations,'¹⁵ organize some groups of bloggers against others. In fact, Egypt is the one national groupings where we see evidence of the kind of large-scale political clustering found in the United States and some other politically active blogospheres (Iran, for instance). Egyptian bloggers comprise nearly one-third of those in our map and form a large structural cluster that contains several attentive clusters, including Secular Reformist, Wider Opposition, Egyptian Youth, Egyptian Islamic, and Muslim Brotherhood.

Secular Reformist

This cluster contains strong political discourse, and consists of bloggers who are critical of Egyptian President Hosni Mubarak and his government. For example, it features a number of bloggers from the loose-knit Kefaya ('Enough') political movement, which is opposed to the president and his plans to install his son as future president of Egypt. Tom Isherwood argues that most influential political bloggers in Egypt got their start as part of the Kefaya

¹⁴ Marc Lynch, "Blogging the New Arab Public." For a broader discussion of Internet development and use in Egypt see also Rasha A. Abdulla, *The Internet in the Arab World: Egypt and Beyond* (New York: Peter Lang, 2007), 45-60.

¹⁵ 'Stop the Imprecations' is a campaign by some, more conservative bloggers, to discourage the use of profanity and impious slang by liberal bloggers.

movement and that the movement, and the 2005 presidential elections in Egypt more generally, propelled Egyptian bloggers from obscurity to the heart of the new Arab public sphere.¹⁶ These bloggers' perspectives are typically secular, and many bloggers have Socialist, some even Marxist, sympathies. This cluster also skews slightly older than other Egyptian sub-clusters, with most of its bloggers (49%) in the 25-35 age range. Some coder text descriptions:

- *This blog is overtly political: At the top it reads in scrolling text: "I blog for freedom. Say no to Mubarak. Say no to Al Ikhwan [the Muslim Brotherhood]." Every entry is related to some political/ social issue - often long essays. The author's viewpoint is secular (though not anti-religious): he heavily criticizes both the government and political Islam, favors democracy and freedom of choice (with regard to the hijab for instance). For him, Muslim society once embraced philosophy and free inquiry, but that heritage has been lost.*
- *This is the blog of an Egyptian member of Kefaya, the secular opposition. In addition to politics, he writes a lot about computer programming, endorsing different kinds of software. He also writes about soccer and the plight of the Palestinians.*
- *This blogger chose to adopt the name of a poet from ancient Egypt, an indication perhaps that his identity harkens back to a pre-Islamic Egypt. The blog is decidedly left-leaning, with the blogger practically eulogizing a figure on the left and a member of the Egyptian Communist party for his dedication to social justice. There is a striking interest in art (film and poetry) that has a political dimension and is engaged with the plight of the downtrodden majority*
- *This is the blog of an Egyptian computer programmer and a Coptic Christian. He writes a lot about the dangers Christians face not just in Egypt, but also around the entire Middle East. He condemns jihadist groups and urges dialogue and understanding.*

Wider Opposition

This cluster is very similar to Secular Reformist in most respects, but contains more bloggers discussing the 'Tomorrow' party, appears to have more expressly leftist and pan-Arab bloggers. It also includes bloggers from other countries writing on similar themes (that are also linking to the same kinds of online resources). It includes some bloggers who deal with AIDS and gay rights.

- *The personal blog of a self identified "Arab Muslim, pan-Arabist, Nasserist revolutionary." The blog posts are mostly about historical events that took place during Nasser's reign, usually during the Six-Day War or the Suez Canal crisis. The blog includes many pictures of memorabilia from that time including postage stamps, currency notes, and newspaper headlines. A smaller portion of posts are political posts on modern events, especially about the Palestinian situation, and at times some local Egyptian protests or strikes.*

¹⁶ Tom Isherwood, "A new direction or more of the same? Political blogging in Egypt," *Arab Media & Society*, Issue 6, (Fall 2008), http://www.arabmediasociety.com/articles/downloads/20080929144105_AMS6_Tom_Isherwood.pdf (accessed May 19, 2009).

- *This is the blog of an Egyptian woman who is a liberal opponent of Mubarak, bitterly criticizing him as a "dictator." She also laments the rise of the Islamist opposition in the form of the Muslim Brotherhood, admitting that it is stronger than liberal opposition parties, but wishing that it were not the case. She also labels the incident in spring 2008, when Hamas blew up the wall with Egypt so that residents from Gaza could buy supplies, "the Hamas occupation of Egypt."*
- *This is an Egyptian opposition blog. It advertises protests, demands the freedom of political prisoners (including both secular leftists and Muslim Brothers), and documents the excesses of the Mubarak government. It calls for tolerance and cooperation among Egyptians of different religious and political backgrounds in order to unite against the government.*
- *This is the blog of a young Egyptian who writes knowledgeably about arts, culture, politics, and social issues. He writes a great deal about the cultural scene in Cairo, but also about everything from Algerian politics to AIDS awareness. His stance on AIDS awareness is particularly interesting in a society that largely considers it taboo.*
- *This is the blog of a gay Jordanian man who writes frankly about his homosexuality, discussing his sexual flings with various men. He writes in support of tolerance for gay people, remaining anonymous for obvious reasons, as such views are very out of line in mainstream Arab discourse.*

Egyptian Youth

Almost half of the bloggers are women, one of the highest percentages of female bloggers in any cluster of the entire Arabic blogosphere.¹⁷ It is also one of the youngest, with half of bloggers in the 18-24 year old range. This group talks more about single life and family life than any other cluster. Among specific political issues discussed in this cluster, support for Palestine is proportionally higher than other clusters. Among cultural issues, bloggers in this cluster discuss poetry, literature, and art often, as well as human rights (by which we mean civil and political rights: freedom of speech, assembly, etc.) Women's issues (rights, status, hijab, feminism, etc.) are much more important to this cluster than to the others (17% compared to a range of 6–9% in the other Egyptian clusters). This cluster also features significantly more bloggers expressing support of Western culture and values than other Egyptian clusters (14% compared to 1–7% in other Egyptian clusters).

The following are descriptions of several blogs indicative of this cluster written by researchers who contributed to this study.

- *This is the blog of an Egyptian who writes a lot about his pride in Egypt's ancient civilization. He also takes on modern politics, criticizing the idea of Egypt moving towards a nuclear program as unhelpful and dangerous, and calling for disarmament throughout the Middle East, including in Iran and Israel.*
- *The blogger writes on a number of different topics, including her emotions, family, politics, and women's issues. She also posts her poetry. All of her writing has a secular and feminist outlook, criticizing radical Islam and the repression of women.*

¹⁷ This estimate is based on the judgments of researchers reading individual blogs and does not include those blogs for which the gender of the blogger was not apparent to the researchers.

- *This is the blog of a young Cairene who writes about film, music, and how much he likes sitting in cafes and drinking coffee. He also touches on politics, criticizing the imprisonment of opposition leader Ayman Nour and the tactics of the security forces.*
- *A young Egyptian's personal blog. Posts vary from very personal ones about her love life and some poetry, to political posts about the government. She seems slightly conservative. She mostly discusses life at her college (she is a teacher's assistant), the political issues they have with the government, and her love life. She is anti-government and dictatorship. She seems to be a central part of an online community, judging from the very large number of comments and amount of discussion.*

Egyptian Islamic

This cluster is focused primarily on the discussion of Islam, similar to bloggers who discuss religion in the Islamic Discourse cluster of the map. These bloggers write primarily about their personal religious thoughts, how Islam fits into their daily lives, how to be better Muslims, and, to a lesser but still significant degree, the Qur'an and Islamic interpretation. Bloggers here occasionally delve into politics and begin to intermingle with the Muslim Brotherhood cluster, which is adjacent on the network map. Some key terms used by these bloggers include general terms related to Egypt, such as Port Said, El Alamein, Adel Hammouda (Editor of the Egyptian newspaper El Fagr), as well as more specific religious terms, Sa'id ibn Jubayr (an Islamic scholar), and religious holidays (Mawlid). Coder comments for this cluster:

- *This blog is largely about religion. The blogger is a devout Muslim, and is well-educated about the intricacies of the faith. He writes some devotional pieces that read like prayers, but also more analytical pieces about Islamic schools of thought and differences between Sunnis and Shi'as.*
- *This blog is about religious issues and choices of behavior, dress, talk, work, etc. It focuses on two things: sanctity of productive work and strong belief in and practice of Islam.*
- *This is the blog of an Egyptian doctor who writes political analysis and about religion. For example, he writes in a non-partisan tone about the recent strikes in the country, discussing what went well while also arguing that change cannot be made through Internet-based activism alone. He also writes about religion a great deal, urging people to read the Qur'an and providing a link for people to listen to readings from it.*
- *This blog is written by a young Egyptian calling himself 'Future Preacher,' and the blog is an entirely spiritual outlet with Qur'anic verses and devotional pieces. It appears to be apolitical, and just devoutly religious.*
- *This is blog of a young Egyptian woman that is devoted to her religious faith. She writes a lot about what she sees as the miracle of the Qur'an. Her site is mostly apolitical in its content, focusing just on preaching, but she also links to Muslim Brotherhood blogs and writes about the suffering of Muslims in Bosnia and Chechnya.*

Muslim Brotherhood

The Muslim Brotherhood has an active online presence, and forms a distinct sub-cluster in the Egyptian part of the map.¹⁸ Despite the fact that the group is technically illegal in Egypt, and that members, including bloggers, are often arrested, members of this cluster are still more likely to blog using their name, rather than anonymously or under an obvious pseudonym. In fact, 78% use their name when blogging, more than any other cluster in the Arabic blogosphere. The Muslim Brotherhood cluster is also more concerned with human rights (41%) than any other group of bloggers on the map. Additionally, 70% of this group is male, one of the higher gender imbalances on the map. This group prefers terms specific to the movement like Ikhwan (brotherhood or Muslim Brotherhood) and Hassan Al Banna (founder of the Muslim Brotherhood), but also terms related to Gaza and the Palestinian conflict.

Researcher comments:

- *A blog run by a member of the Muslim Brotherhood that basically documents and publishes news about Muslim Brotherhood detainees. It is very anti-government and anti-Mubarak, and focuses mostly on the detainment of many leaders of the Brotherhood in mid-2007 and the use of military tribunals as opposed to civil courts in their cases.*
- *This is the blog of a prominent Muslim Brotherhood activist. He champions the Brotherhood's cause in Egypt, but also uses his blog to challenge the Brotherhood leadership, urging them to take a more politically active stance and reach out to the secular opposition.*
- *Blogger discusses religion, religious issues, and beliefs, mixed with philosophy. This blogger posts links to Muslim converts from different countries in North America and Europe.*
- *The blogger is very interested in how Islamist movements can spread influence culturally and resist Western missionary activity, which he sees as being inseparable from Western imperial ambitions. Feminist thinking and so forth are viewed as Western ideologies intended to capture minds and destroy Islam from within.*

Levantine/English Bridge

This region features a combination of bloggers from the countries of the eastern Mediterranean referred to historically as the Levant, including Lebanese, Jordanians, and Palestinians. They are joined in this section of the network, which connects to the United States and international blogosphere, by 'bridge bloggers' from other countries, who write in English. There is no hard division in the network between English and Arabic blogs, and

¹⁸ For a discussion of the Muslim Brotherhood's online activities see Pete Ajemian, "The Islamist opposition online in Egypt and Jordan," *Arab Media & Society*, Issue 4 (Winter 2008), http://www.arabmediasociety.com/topics/index.php?t_article=183 (accessed May 19, 2009) as well as Marc Lynch, "Young Brothers in Cyberspace," *Middle East Report* #245, Winter 2007, available at <http://www.merip.org/mer/mer245/lynch.html> (accessed May 20, 2009). For a general overview of the Muslim Brotherhood, its transformation over time, and misunderstanding in the West see Juan Cole, *Engaging the Muslim World* (New York: Palgrave Macmillan, 2009), 41-82, and Mona El-Ghobashy, "The Metamorphosis of the Egyptian Muslim Brothers," *International Journal of Middle East Studies* 37/3 (August 2005).

many of the bloggers mix the two languages in various proportions. The region includes several different attentive clusters, including English Bridge, Jordan, and Lebanon. The English Bridge area also includes most Iraqi bloggers, who appear here instead of in their own national cluster, perhaps because their dense connections with US think tanks, journalists, and partisan political bloggers overwhelm those that would arise from serving a domestic Iraqi audience. This is also where the international bridge blogging network Global Voices is found.

English Bridge

While still consisting mainly of bloggers living in the Middle East, this sub-cluster has more expatriates than most (16.7%). They appear to be tied together more by their linguistic choices than any demographic or national connections. However, they are very political and discuss international news as much as local affairs, with strong attention to human rights, women's rights, and Western culture and values, which they mainly support. They are also the most active critics of the United States of any cluster on the map (18.8%), though they are one of its biggest defenders as well (7.1%). A number of American blogs link to this bridge cluster, which includes Iraqi bloggers who write in English, as well as bridge blogging communities like Global Voices and MidEastYouth.com. Because they are the main link between the Arabic and international blogospheres, perhaps it is to be expected that we see the most discussion of issues of concern to the West within the bridge.

Terms used by this group include: place names in Iraq (Basra), Amnesty International, the United States, United Kingdom, Bahrain, Oman, and women's rights. Key outlinks include English language news sites (New York Times, BBC, The Guardian, Al Jazeera English) and several major Iraqi bloggers.

Researcher comments:

- *A group blog by young writers who view themselves as avant garde. One entry deals with walking in urban space for women, the kinds of encounters that produces, and what that says about the political culture in contemporary Egypt. The site is highly secular, very critical of those who view Islam as the solution for the current quandary, and highly skeptical about the compatibility of this kind of thinking with democracy.*
- *The blogger is a Syrian dissident who seems quite aware of what is happening in American politics and in Middle Eastern struggles for regional power, and he is vehemently opposed to the Syrian ruling family. He is a strong advocate for political reform within Syria and would be happy to see the Syrian regime go.... He seems attuned to the policies of the neo-cons, in the sense that he believes democracy must come to the region, though he doesn't seem terribly enamored of the intelligence of those advocating such policies that merely strengthen Syria, Iran, and Hezbollah.*
- *This blogger seems to have much experience with traveling around the world and now is settled in Bahrain, her home country. She is critical and somewhat sarcastic of the role of the Arabic language, especially in education, and recently discusses how leaders in her country consider everything in Arabic to be related to Islam. She also discusses Western poetry, literature, and film, and the complexity of living in Bahrain;*

- *The blog includes discussions and news about the current situation in Iraq: Five years after the United States occupied Iraq with the stated goal of bringing democracy to its people, the war has essentially ruined the country's economy, society, and sovereignty;*
- *The blog is in English and deals with a Sudanese's thoughts and ideas on international and national issues.*
- *This is the blog of a Syrian studying in Japan who writes mostly in English with a little Arabic. He writes fondly about his childhood in Lattakia, adjusting to Japanese culture, the global presence of Islam, and campaigns for the freedom of a Syrian blogger who was jailed for posting comments that the regime found objectionable.*

Lebanon

Lebanese bloggers form a distinctive sub-cluster within the broader Levant/Bridge region of the map.¹⁹ This group writes in English, Arabic, and occasionally French. Although a minority, this cluster also includes bloggers from Palestine. Topics range from political issues in both Lebanon and Palestine to photography, poetry, medicine, and personal, diary-type reflections. Domestic leaders are widely criticized (51.6%), but also supported (29.0%), indicating a robust attention to national politics. Religion appears to be discussed less often, almost exclusively in terms of personal thoughts and reflections. Western culture and value are discussed widely, supported and criticized in equal measure (~32% each). Key online resources for this group include Lebanon-focused news sites, blogs, and political party sites (Free Democratic Movement Party), in Arabic and English. Terms preferred by this group reveal an interest in international, domestic, and Palestinian policy: military, America, Iran, Israel, Palestine, and Obama.

Researcher comments:

- *This is the blog of young Lebanese woman, who is a Christian and a supporter of the Free Patriotic Movement, the Lebanese Christian faction that is allied with Hezbollah. She jokes that she supported the "evil Syrian-Iranian axis" by voting for the party. She defends the nationalist credentials of the movement, and devotes many posts to justifying her support for the Hezbollah-aligned movement as a Christian.*
- *A political blog. Keep the voices of the Lebanese and those who visited Lebanon recently loud and clear. NO TO WAR AGAIN. NO TO OCCUPATION. NO TO TERRORISM. For the land of milk and honey, Lebanese Youth Call on US to Respect Democracy.*
- *This is the blog of a Palestinian living in France. He writes mostly about the Palestinians, decrying their conditions in refugee camps, in exile, and under occupation. He also writes a little about the situation of Arabs and minorities in France.*

¹⁹ For discussion of Lebanese political blogs see Sune Haugbolle, "From A-list to webtifadas: Developments in the Lebanese blogosphere 2005-2006," *Arab Media & Society*, 1, Spring 2007, <http://www.arabmediasociety.com/?article=40>, accessed May 19, 2009.

- *A blog of a Lebanese doctor. All of his posts are exclusively self-authored poems. Many of his poems are nature poems. He uses the family and romantic relationships as metaphors for his relationship with nature. Other poems are about love.*
- *The blog of a young Lebanese woman who writes about looking for love and trying to find the right man. She also posts a lot about herself, expressing a great deal of self-confidence, but also complaining of daily frustrations.*
- *This Lebanese blogger works as a translator, and keeps a mainly personal blog. One post also deals with the time when downtown Beirut was closed off due to a long demonstration - however, once again, the blogger simply discusses the personal experience (i.e. not being able to go to work, etc.) rather than taking any political stand.*

Jordan

This cluster contains mainly Jordanians, but also some Palestinians. It has the highest proportion of women in the Levantine/English Bridge group. Female bloggers comprise 40.2%, 30.2%, and 33% of the Jordanian, Lebanese, and English Bridge clusters, respectively. Women's issues are more widely talked about here than any other cluster (54.2% vs. 19.1% across the map). Jordanian bloggers are less critical of domestic leaders than Lebanese (29.2% vs. 51.6%), probably reflecting Lebanon's more democratic system. Bloggers in this cluster use the word 'Palestine' more frequently than any other cluster, and 'Nakba' (1948 Palestinian exodus) more frequently than all but the Egyptian Religious cluster.

Researcher comments:

- *This blogger is critical of how political activism is conducted (related to Gaza, for example) and wishes it was more focused and directed. Like many secular Arabs, she seems frustrated by the narrow range of options available. The Islamists are passionate and highly organized, and other options, tribal and secular forces, are self-absorbed and somewhat indifferent to politics.*
- *The blogger sees herself as combination of Muslim and traditional Circassian.²⁰ Her blog includes English and Arabic entries. She expresses her support for the Ossetians, whom she feels have suffered unjustly at the hands of the Georgians.*
- *This is the blog of an Arab-Israeli girl, a Palestinian who is living in Jerusalem where she is a student. She writes a bit about her life, but mostly about her country and her Palestinian identity, focusing a lot on historical injustices and indignities suffered by her people.*
- *This blog, by a Jordanian, is almost entirely in English with just a few items in Arabic. He writes mostly about Palestine, commemorating the Nakba and condemning the Israeli occupation. He also writes about soccer.*

²⁰ A minority group, originally from the Caucasus, many of whom settled in the Levant in the 19th Century.

- *This is the blog of a Syrian living in Jordan who writes humorous posts about his life, which he considers to be absurd. On more serious notes, he promotes his businesses in Jordan and writes about life back in his native country. He writes in English more frequently than Arabic.*

Maghreb/French Bridge

The Maghreb is a term used to refer to the region of North Africa that includes Morocco, Algeria, and Tunisia.²¹ This cluster of the network contains North African bloggers, many with secular outlooks, using combinations of French and Arabic, and also bridges to French-language sources. Another cluster, described later, contains a number of more religious Maghrebi bloggers writing exclusively in Arabic. The Maghreb/French Bridge cluster is overwhelmingly male (almost 90%). In terms of political topics, this cluster is similar to the map at large: critical of domestic leaders, terrorism, and foreign leaders, while supportive of Palestine. Among the discussions that focus on topics of culture and values, this cluster calls attention to human rights more than any other topic, although poetry, art, and literature are also important. It also features slightly older bloggers on average than those of other clusters. It is also one of the few clusters with more anonymous authors than those using their name (51.3% to 48.7%), and contains a relatively high proportion of expatriates (17.2%).

Researcher comments:

- *This (Tunisian) blogger is very interested in the collective fate of countries in the region. He critiques what he calls a culture of fear that seems to be paralyzing it. He sees this fear holding the culture back and traces it to various aspects of life, including child rearing. One might say he is making a plea for courage and a gradual rebuilding of self-esteem for a culture that has lost its sense of self. The blogger also deals with the polarization that is beginning to occur between the faithful and the secular, making a plea for the modernization of Sharia.*
- *This is a blog by a middle-aged Tunisian woman who writes about striking a balance between Islam and civil society. She argues passionately in defense of her Islamic faith and makes it clear that she works to follow its laws, but she nonetheless calls for a civil state with a separation between civil law and religious law.*
- *This is the blog of a Tunisian who writes about his personal life and being in love, but also about more politically-charged topics. He rails against the legacy of colonialism and European political influence in North Africa, and also writes about domestic politics (especially as it relates to the Internet and press freedom) in Tunisia.*
- *This blogger is a Tunisian man living in Belgium who writes about personal topics often, sometimes writing open letters to his father. He also writes some about the experience of being a Muslim in Europe, defending his faith in what he feels is an Islamophobic climate.*

²¹ For a discussion of blogging in Morocco see Aziz Douai, "Offline Politics in the Arabic blogosphere: Trends and Prospects in Morocco," in *International Blogging: Identity, Politics, and Networked Publics*, eds. Adrienne Russell and Nabil Echchaibi (New York: Peter Lang, 2009) 133-149. The plight of Tunisian bloggers, which rarely discuss political issues, is also discussed by Marc Lynch, "Blogging the New Arab Public."

- *This is the blog of a Tunisian man who is critical of police brutality in the country and of the government in general. He also takes a very liberal stance on issues, and condemns Islamist movements in North Africa, accusing them of brainwashing young people.*
- *This blogger is militantly secular. He seems to have been actively involved in a global day against the hijab (veil), which he sees as being representative of backward Salafi thinking, including the myriad dangers of takfir (calling violence upon apostates).*

Saudi Arabia

This cluster devotes far less attention to domestic political leaders than other clusters, both in terms of support and criticism.²² This group writes primarily in a personal diary format, even more than other clusters. Automated term frequency analysis shows a preference for technology and fewer mentions of political terms or personalities.²³ At 46%, this cluster has more female bloggers than most other clusters. While still a small minority, it also has more expatriates than other clusters. There were several posts that argued for the release of a jailed Saudi blogger, Fouad al-Farhan, who was arrested for the writing on his blog but eventually released.

Researcher comments:

- *Identifying himself as 'The Philosopher,' this blogger takes on political issues concerning the Arab world as a whole. He is critical of his native Saudi government, including for its recent imprisonment of a fellow Saudi blogger. He also condemns the American and Israeli positions in the Middle East.*
- *A personal blog of a Western Saudi Arabian college student. She blogs about her life at the college and her interests, but not much beyond her daily activities and problems. She has some insightful posters on the blog. One is against drugs, while another protests against Arabic Wikipedia articles defaming the prophet. Also, she has one rare political poster that is very critical of Hezbollah.*
- *This is the blog of a young Saudi woman who writes about her thoughts, friends, dreams, and her prospects for a university education. She also writes Islam and her faith in God.*
- *Blogger is either attempting to expose the status of women in Saudi Arabia or just seems to be concerned with her own state of happiness.*
- *This is a personal blog, for the most part on religious subjects and issues dealing with the challenges facing the Arab and Muslim world. This blogger also discusses purely personal issues as well as enthusiasm for computers and technology.*

²² This finding on the Saudi blogosphere is supported by Marc Lynch when he writes, "In Egypt, Bahrain, and Kuwait, for example, bloggers have played a key role in mobilizing contentious politics. Other Arabic blogospheres, such as Jordan and Saudi Arabia, remain more on the political sidelines, although they have increasingly taken note of the exploits of their activist peers." Marc Lynch, "Blogging the New Arab Public."

²³ Top terms relative to other clusters include: technology, WordPress, computer programming, Flickr, Twitter, and Microsoft.

- *This is the blog of a Saudi woman who writes mostly about cultural issues, including Islam and modernity and relations between the sexes in the kingdom. She also defends Fouad Al Farhan, a political blogger who had been jailed by Saudi authorities.*
- *This is the blog of a young Saudi economics student who writes almost exclusively about new computer software, his apparent passion.*
- *This is the blog of a Saudi, originally from Jiddah, studying in Tampa, Florida. He writes mostly about technology, hawking new computer and mobile phone products. He also writes about films, his faith in Islam, and speaks out in support of bloggers' rights in his native Saudi Arabia.*
- *This is the blog of a Saudi man who writes about his life and about politics. He writes in support of the Palestinians and against the US occupation of Iraq, and he also complains about some domestic issues, such as what he considers to be poor medical services available in Riyadh.*

Syria

Bloggers in this group are located almost exclusively in Syria and write primarily about domestic issues, including politics. Syrians are among the least likely to express support for domestic political leaders compared to other clusters. This cluster's discussion of religion is dedicated predominately to personal religious thoughts and experiences, as well as religious poetry and Sunni Islam. This cluster has the second largest concentration of known male bloggers on the map, at 87%, while only 13% are female. Half of all bloggers in this cluster are in the 25-35 age range, making it slightly older than other clusters and the blogosphere on whole, but not by much. Coders found that this group was more critical of western culture and values than other clusters and one of the least likely to discuss movies and television.

Below are coder comments about blogs in this cluster:

- *This is the blog of a young Syrian who is interested in politics and technology. He writes mostly about world and regional politics, criticizing the United States, the House of Saud, and Israeli policy in Gaza. However, he also criticizes the Syrian government's decision to put a blogger in jail, which reflects both the way in which political blogging is dangerous in Syria, but also the way blogging is one of the few avenues to offer any criticism of the regime in the country.*
- *This is the blog of a Syrian student who writes lovingly about Damascus. He devotes much of his blog to writing about life in the city, his studies, and his support for the Palestinian cause.*
- *This is the blog of a Syrian liberal dissident now living in the United States. He urges the release of Syrian dissidents in jails, call for democracy and reform in Syria, and supports secular liberal around the Arab world. He is also very critical of the Israel and, to a lesser extent, the United States, considering them obstacles to Arab democracy.*

- *This is the blog of a young Syrian man who writes about his daily emotions, including joy and sadness. His blog is mostly personal and not political, but he asks what the effects of the struggles in Palestine and Iraq have on Arab culture and group consciousness.*
- *This is the blog of a Syrian student who writes personal things about his own life but also political pieces. On politics, he goes about as far as a blogger in Syria can go without getting arrested, denouncing corruption in Syria and calling for greater Internet freedom and less censorship. He never directly takes on President Assad or specific government figures, which is probably what keeps him out of jail.*

Kuwait²⁴

This cluster consists of Kuwaiti bloggers, and is split between two sub-clusters, one that prefers English and the other that prefers Arabic.

Kuwaiti English

Bloggers in this sub-cluster prefer English language sources and also often write in English. This group includes a number of Kuwaiti expatriates as well as foreigners living in Kuwait who write in English. They are more likely to write anonymously than to use their name, and are more likely to be male than female compared to Arabic bloggers. This group is also older than their Arab counterparts, with 77% of them in the 25-35 year old age range. They discuss international news and policy, economic issues, ethnic minority issues, and women's issues more often than Kuwaiti Arabic bloggers. This group also talks more about family life and single life than their Arabic counterparts in Kuwait. This group is also more likely to be supportive of Western culture and values and more inclined to write about technology, human rights, poetry, literature, and art than Arabic speaking Kuwaitis. Kuwaiti English bloggers are also much more likely to be critical of terrorism, less likely to support Palestine, less likely to express support for domestic political leaders, and more likely to criticize foreign political leaders than their Arabic counterparts.

Here are some representative coder descriptions of blogs in this sub-cluster:

- *This blog is written by three liberal, pro-democracy Kuwaitis. They are all young men with an average age of about 26, and they post items about protests and the struggle for reform in the monarchy. They are critical of the government, and secular in their outlook.*

²⁴ Kuwaiti bloggers are among the more politically active in the region according to Marc Lynch, and gained an audience during the 2006 succession crisis following the death of the Emir. Bloggers then entered the debate on reducing the number of electoral districts to cut back on corrupt electoral practices and jumped into the Parliamentary elections with a vengeance in 2006. Marc Lynch, "Blogging the New Arab Public," see also Mary Ann Tetrault, "Kuwait Annus Mirabilis" September 7, 2006 <http://www.merip.org/mero/mero090706.html>. For an alternative view see Deborah Wheeler, who argues that political use of the Internet in Kuwait is constrained, economic use is latent, and the Internet will not necessarily lead to enhancements of personal freedom and power in Kuwait: Deborah L. Wheeler, *The Internet in the Middle East: Global Expectations and Local Imaginations in Kuwait* (Albany: State University of New York Press, 2006) 61-103.

- *This is the blog of a Kuwaiti with a philosophical bent who writes poetry and long analytical posts that address complex topics like keeping Kuwait's Muslim identity in an age of globalization. He also writes about the country's partial democratization process.*
- *This is the blog of a Kuwaiti who writes about historical topics, such as old books he finds. He also writes a little about Arab politics and art.*
- *The blogger is focused on Kuwaiti politics, domestic issues, and public affairs, sometimes using sarcasm and jokes to comment on internal affairs.*
- *The blog is in English and deals with women's issues regarding family life, pregnancy, knitting, taking care of the baby, etc.*

Kuwaiti Arabic

This group of bloggers primarily writes in Arabic and prefers Arabic language sources. They are located predominately inside Kuwait and are more likely than Kuwaiti English bloggers to use their name when writing, although the majority of Kuwaiti bloggers in both clusters write anonymously. This sub-cluster is younger and has more females than their English counterparts, but the majorities in both sub-clusters are male. In terms of topical issues, Kuwaiti Arabic bloggers are far more likely to discuss domestic news and Islam (including the Qur'an, theology, history, and Sunni thought and practice) than Kuwaiti English bloggers. Automated term frequency analysis shows that Kuwaiti Arabic bloggers are much more likely than their English speaking counterparts to discuss political issues, and indeed use political names and terms more than most other clusters in the Arabic blogosphere. Nearly all top terms were related to top Kuwaiti politicians or those that figure prominently in recent political debates.²⁵ Coding results reveal that among specific political topics this group is more likely to support domestic political leaders, more likely to be critical of Israel, and more likely to support Palestine than the Kuwaiti English sub-cluster. This group also talks more about pop culture topics (movies, TV, and music) than their English counterparts.

Coder comments from the Kuwaiti Arabic cluster:

- *A blog about Kuwait domestic affairs and increasing tensions between Islamists and Liberals.*
- *A Kuwaiti liberal who blasts corruption in the monarchy and urges people to vote for reformers in parliamentary elections. He also celebrates the execution of Saddam Hussein.*

²⁵ Top terms (individuals) included: Faisal Al-Muslim Al-Otaib, Saleh Al-Mulla, Ali Al-Rashid and Mohammed Hayef Al-Mutairi.

- *A Kuwaiti liberal who is critical of his government but seeks to improve it and reform it from within. He highlights news stories about problems in Kuwait and mistakes the government has made, writing that he hopes to see things improve.*
- *This blogger writes about Kuwaiti politics, chronicling the push for liberalization and reforms in the country. She is clearly supportive of reform, with banners on her blog calling for greater free speech and constitutional rights. She is also somewhat critical of the West.*
- *This is the blog of a Kuwaiti woman which is interesting for her two main focuses-- Islam and men. She sees no apparent contradiction between her faith, which she writes frequently about, and her desire to meet men and date.*
- *This is the blog of a Kuwaiti who posts random thoughts about God and religion. He is a devout Muslim and many of posts simply praise God and Muhammad.*
- *This is the blog of an educated Kuwaiti man who writes about the effect of high oil prices on the kingdom's economy, as well as legislation in the parliament and the slow reform movement towards more democracy.*
- *This is the blog of a Kuwaiti woman that mixes the personal and political, with posts about her search for love interspersed with posts complaining about rising bread prices in the monarchy and criticism of Shi'a Kuwaitis who use supporting Hezbollah as a cloak for sectarianism, urging an end to sectarian agitation in Kuwait.*

Islam Focus

In the corner of the map opposite the English Bridge, we find a region with several attentive clusters of religious bloggers focused on Islam. This region connects more closely with the religious side of the Egyptian blogosphere, and with the Saudi cluster. Two attentive sub-clusters stand out, Islamic Discourse, which prioritizes religious interpretation, and Arabic Maghreb, which contains bloggers who are mainly from the Maghreb (along with some from Egypt, including Muslim Brotherhood members) who write entirely in Arabic, sometimes criticizing those who write in French. A lot of bloggers in this region use the Maktoob blog service, which seems popular with more religious bloggers.

Islamic Discourse

This group is focused primarily on discussion of Islam, including interpretation, personal religious thoughts, and daily application of Islam, teachings, fatwas and so forth.²⁶ Sixty-eight percent of bloggers in this cluster discuss the more theological aspects of Islam, the highest proportion in the map (the second highest cluster, Egyptian Islamic, is still only

²⁶ There are a number of sources that discuss the use of the Internet for Islamic discourse and interpretation including: Gary Bunt, *iMuslims: Rewiring the House of Islam* (Chapel Hill: University of North Carolina Press, 2009); Gary Bunt, *Islam in the Digital Age: E-jihad, Online Fatwas and Cyber Islamic Environments* (London: Pluto Press, 2003); Eugenia Siapera, "Theorizing the Muslim Blogosphere: Blogs, Rationality, Publicness, and Individuality," in *International Blogging: Identity, Politics, and Networked Publics*, eds. Adrienne Russell and Nabil Echchaibi (New York: Peter Lang, 2009) 29-46; and Jon Anderson, "The Internet and Islam's New Interpreters," in *New Media in the Muslim World*, eds. Dale Eickelman and Jon Anderson (Bloomington: Indiana University Press, 2003) 45-60.

24.1%). This is a transnational group with bloggers from Egypt, Saudi Arabia, North African, and Gulf countries, though Egyptians are the biggest group. It is also predominately male; approximately two-thirds of bloggers identifiable by gender are men. This group is much more focused on discussion of Islam than on politics, with much less advocacy of political reform or discussion of political leaders than most other clusters. This cluster's interpretation of Islam also appears to be more traditional or conservative, and criticism of other faiths is very high compared with other clusters (27.3% vs. 3% average for the network overall). Terms that this cluster prefers are overwhelming related to Islam, Islamic figures and thinkers, and other religions.

Researcher comments:

- *Personal Egyptian (Islamic) blog. The title or premise of the blog is 'Have you forgotten Heaven?' with a picture of a female Hamas supporter holding up a Qur'an. It is a very Islamic blog with a newsreel at the top with news on Islamic 'resistance' (Lebanon, Sudan, Somalia.) Posts could be split into three categories. First, posts that hint towards the creation of an Islamic state/caliphate with many of them being historical, about the downfall of the Caliphate and other Islamic nations, or about Islamic movements around the world. Secondly, a lot about the "other's" (I can only guess she means non-Muslims) and treatment of Muslims in the West and other places, including being jailed and tortured. Thirdly, posts about how to be a good Muslim. Many posts are also about "Muslim" Palestine. The blogger is Sunni and seems to be Salafist/Wahabbist, though she protests these labels in one of her posts.*
- *This is an Egyptian blog devoted to spreading Islam. The blogger praises the virtue of the religion, and links to several sites with information and videos. The blogger also descends into sectarian territory, blasting Christianity and Judaism and singling out Egypt's Coptic Christians for criticism.*
- *The blogger has an attraction to the televangelist Amr Khaled and is devoted to this new type of preacher who makes Islam more user friendly to those who have not yet reached a high level of faith or Iman. The blogger recognizes the importance of using media, including blogs, to spread faith amongst the young and seems to have a somewhat paranoid vision of how foreigners (Israelis and Americans) are fighting Islam. On the other hand, she cites a story about an American woman who started out with the intention of subverting Islam only to convert.*
- *This blogger is vehemently against secularism, viewing the separation of church and state as incompatible with Islam. It is this very incompatibility that distinguishes Islam from Christianity. Blogger seems closed to other cultural influences.*
- *The blogger is very hostile towards the Coptic minority in Egypt, citing their monopoly over the economy. He also seems to have entries that have to do with Muslims (which are termed dogs) who have become apostates.*
- *This is a religious blog with theological information about Islam, and various fatwas on different occasions and issues, such as Valentine's Day.*

Arabic Maghreb

These bloggers are mainly Magrebis who only use Arabic, and are in general more religious than their Francophone counterparts in the Magreb/French Bridge cluster. There are also some Muslim Brotherhood bloggers from Egypt who, because of similar linking preferences, are in this cluster as well. Some of these Muslim Brotherhood members appear to be young and particularly conservative (perhaps even militant), in opposition to the more numerous young, reform-minded brothers we see in the Muslim Brotherhood attentive cluster. The analysis of word and term frequencies shows this cluster to be particularly attuned to Gaza and Palestine, including Hamas leaders and militant organizations such as the Hamas' Izz ad-Din al-Qassam Brigades and the Popular Resistance Committees. They are less interested in theology, and more in politics, than the Islamic Discourse cluster.

Researcher comments:

- *There is an entry in which the blogger discusses the possibility of being an Islamist and of political dissent vis-a-vis the ruler, given the Qur'anic verse to obey the wali. What he states, quite clearly, is that this form of opposition can take on a military form (using weapons to be exact), so I am concluding that he implicitly agrees that violent means can be used to oppose a corrupt ruler. The blogger is very interested in how Islamist movements can spread culturally and resist Western missionary activity, which he sees as being inseparable from Western imperial ambitions. For him minorities are quite suspect. The Bahai's, for instance, are seen to thrive wherever Zionism flourishes. The Ahmadiyahs are likewise seen to be puppets of British colonialists. In keeping with this logic, feminist thinking and so forth are viewed as Western ideologies intended to capture minds and destroy Islam from within.*
- *This is the blog of a Moroccan woman who writes about her life as a student, her opinion that Moroccans ought to always speak in Arabic instead of French, and has posts about her daily life, including one in which she became upset over misplacing a large sum of money.*
- *Blogger is critical of the domestic situation in Morocco and rampant corruption in the government. Blogger condemns the political situation, corruption, and false statements in government project implementation.*
- *The blogger is advocating 'Without French,' a campaign dedicated to promoting the use of Arabic in Morocco and protesting the widespread use of French as a main language. Most posts are about preserving Moroccan identity (both Arab and Amghazi/Berber) including news about the issue in the surrounding areas.*
- *This is a very conservative Muslim Brotherhood blogger. While some members of the group focus more on politics than religion and reach out to moderates, this blogger writes mostly about religious devotion and describes himself as "a soldier in the Muslim Brotherhood army."*

Baha'i

This small cluster is made up of Baha'i bloggers or bloggers that discuss Baha'i topics. Among religious topics, this group talks overwhelmingly about other religions, including criticism of other faiths, but also interfaith dialog. Since the Baha'i are persecuted in some countries, this may help explain why this group is more likely to blog anonymously (75%)

than any other cluster. This group of bloggers is also the most critical of political Islam of all clusters. They also seem to talk about family life more than other clusters.

Coder comments:

- *The blog exists to promote the Baha'i faith, encourage members of the Baha'i religious community, and to tackle spiritual questions such as why God allows there to be so much pain in the world.*
- *This blogger reveals nothing about his/herself, and simply reposts news articles and information about the status of Baha'is in Egypt, criticizing the treatment that the tiny religious minority group suffers.*
- *This is the blog of a grandfather and grandmother who are members of the Baha'i faith. They write about their kids and grandkids, and also about their religion. They focus especially on the troubles faced by their co-religionists in many nations in the Middle East.*
- *This is the blog of a young Baha'i man in Egypt, who promotes his faith and urges tolerance for his co-religionists. He focuses in particular on the recent identity card controversy in Egypt, complaining about the recent government ruling that disallows claiming Baha'i faith on one's national identity card, leaving Baha'is a choice of listing another religion or being unable to work legally in Egypt and receive social benefits.*

RESULTS: HUMAN CODING

To complement automated techniques, we also conducted exploratory human coding to understand what issues bloggers discuss, their opinions on different topics, and demographic data about the coders. A team of Arabic speakers coded 4,370 Arabic language blogs from April 2008 to March 2009. Two code sheets, or questionnaires, were used to determine demographic data about each blogger (age, location, gender, etc.), as well as topics that bloggers discuss. Coders were able to add topics not listed in the code sheet, and, finally, coders were required to write a short description of each blog or blogger and what was unique or stood out about them. These researchers coded blogs 'live,' meaning they read each blog online at the time the blog was coded. They were instructed to code based on the most recent five to seven recent posts, and did not read archived material. Thus, these results indicate a 'slice of time,' and so it is a guide to the relative salience of particular concerns, rather than a comprehensive assessment of what topics bloggers may have discussed over a long period.

Topical and Issue-Based Coding Results

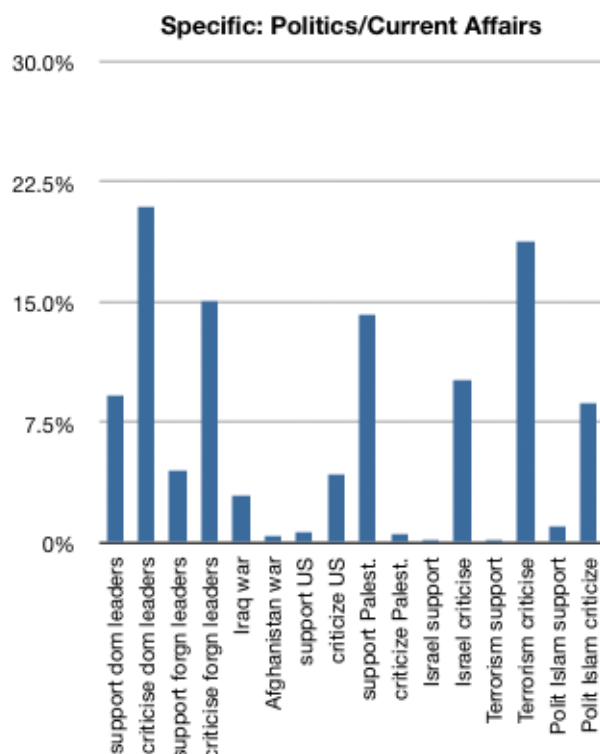


Figure 2: Specific Political Topics

We asked coders to indicate the specific political issues that bloggers discussed. This included whether bloggers showed support or criticism for foreign and domestic leaders, the United States, the wars in Iraq and Afghanistan, Israel, Palestine, terrorism, and political Islam.

Out of the over 4,000 blogs that we coded, less than 1% express explicit support for terrorism. Further, 19% of bloggers are explicitly critical of terrorism, which makes it the second most popular specific topic we asked about. Also, only 1% show support for political Islam while 9% criticize it. These results, combined with other findings throughout the study, seem to indicate that concerns of policy makers in the United States regarding the use of the Internet for spreading hate and support of terrorism may be overblown.²⁷ A number of recent polls of public opinion in the Arab world support the finding of low, and decreasing, support for violent extremism. According to the Pew Global Attitudes Project's report 'Global Public Opinion in the Bush Years (2001-2008),' although mistrust among Muslims and the West remains, there was a notable decline in support for suicide bombings in Muslim public opinion from 2002 to 2008. In Lebanon, support for suicide bombings dropped from 74% in 2002 to 32% in 2008 and in Pakistan from 33% to just 5%. Further,

²⁷ However, as we note in the discussion section of this paper, terrorism means different things to different people, needs to be put in context, and additional research is needed in this area. Here we are referring to anti-Western extremist groups such as Al-Qaeda who carry out attacks on civilians for political ends.

Pew found that support for Osama bin Laden has plummeted in recent years in Muslim countries, dropping from 56% in 2003 to 19% in 2008 among Jordanian Muslims, while just 2% of Muslims in Lebanon and 3% in Turkey were confident that Osama bin Laden would do the right thing in world affairs.²⁸

At least in the kind of online ‘public publics’ represented by Arabic blogs, discourse appears varied, thoughtful, and for the most part moderate. This is not to say that anti-Western ideas are absent, or that groups like Hamas and Hezbollah do not have significant support, but that these ideas are countered by others, and support of Al Qaeda and civilian attacks is very rare when measured against the number of voices concerned with more constructive ideas and compassionate attitudes. To the extent that pro-terror messages thrive online, they do so in the more shadowy ‘private publics’ constituted within ephemeral chat groups, bulletin boards, and closed social networks.

Additionally, bloggers are far more concerned with domestic political issues than with the United States or its wars in the region. Criticism of domestic political leaders is the most frequent political topic of those we coded, followed by criticism of terrorism. Criticism of foreign leaders is also significant, as is support for Palestine and criticism of Israel. Regarding concerns in American public diplomacy circles that the Arabic blogosphere and media are tilting Middle East audiences against the United States and its policies, the United States and the wars in Iraq and Afghanistan are discussed far less than other topics. Interestingly, most discussion of the United States is found in the Levantine/English Bridge, the part of the network which connects with a wider international, including American, audience. In the English Bridge cluster, 18% are critical of the United States, while 7.1% express support. Overall across the network, 4% are critical of the United States and 1% expresses support. Only 3% discussed the war in Iraq, and only 6 blogs (less than 1%) discussed the war in Afghanistan. These results seem to indicate that Arabic bloggers, as we found with Iranian bloggers previously, are far more concerned with issues at home and with their own political leaders than they are with the United States.

²⁸ Andrew Kohut, Richard Wilke, Erin Carriere-Kretschmer and Kathleen Holwart, “Global Public Opinion in the Bush Years (2001-2008),” Pew Global Attitudes Project, December 18, 2008, available at <http://pewglobal.org/reports/display.php?ReportID=263> (accessed May 27, 2009). See also Steven Kull, et al, “Public Opinion in the Islamic World on Terrorism, al Qaeda, and US Policies,” WorldPublicOpinion.org, Program on International Policy Attitudes, U. of Maryland, February 25, 2009, available at http://www.worldpublicopinion.org/pipa/pdf/feb09/STARTII_Feb09_rpt.pdf (accessed May 28, 2009), which found large majorities denounced attacks on American civilians and that views of Al Qaeda were mixed, with majorities supporting their ‘Islamist goals’ but disagreeing with their tactics of attacks on civilians; Shibley Telhani, “2009 Annual Arab Public Opinion Survey,” University of Maryland/Zogby International, available at http://www.brookings.edu/events/2009/~media/Files/events/2009/0519_arab_opinion/2009_arab_public_opinion_poll.pdf (accessed May 27, 2009), which found that 32% did not empathize at all with Al Qaeda, up from 21% in 2008, and that only 7% supported their methods of operation. The Human Security Project’s “Human Security Brief 2007,” available at <http://www.humansecuritybrief.info/access.html> (accessed May 28, 2009) review of public opinion polls also found a trend of decreasing support for terrorism and al-Qaeda in the Muslim world. For a criticism of extrapolating Arab public opinion from surveys in the region and the overemphasis on questions about perceptions of the United States see Kenneth Pollack, “Slippery Polls: Uses And Abuses of Opinion Surveys from Arab States,” Washington Institute for Near East Policy, April 2008, available at <http://www.washingtoninstitute.org/templateC04.php?CID=290> (accessed May 28, 2009); Pollack also notes the underreported but dramatic rise in opposition to anti-American terrorism in the polls.

Coders were also asked to identify general topics discussed by bloggers. Coders could choose as many topics as applied from an initial list, and could add additional topics. Topics were broken out across three broad categories: 1) politics and public life, 2) culture, and 3) personal life. Among general politics and public life topics (Figure 3), it is clear that domestic news (64%) and international news (46%) are the most popular, which seems to support the theory that news is an important driver of content in blogs. Also popular with bloggers, but at lower levels, are ethnic minority issues (25%) and economic issues (21%). Even fewer, 19%, discussed either domestic political reform (advocating change within the current regime or system) or women's issues (rights, status, hijab, feminism, etc.).

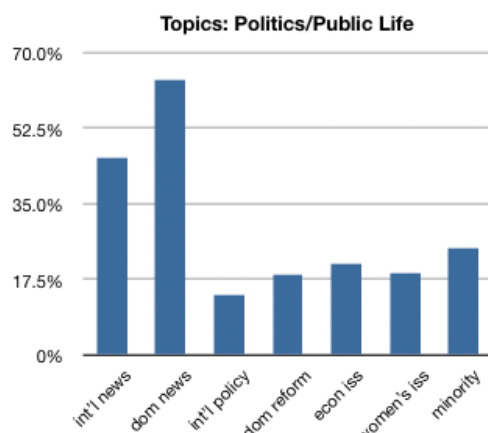


Figure 3: Politics and Public Life

Religion is also a popular topic in the Arabic blogosphere. Religions other than Islam are rarely discussed (4%), and there is very little criticism of other faiths (3%). Instead, Arabic bloggers mainly write about their own personal religious thoughts and experiences (35%). Fewer, though still a substantial number (12%) write about Islam generally, including the Qur'an, theology, Islamic interpretation, and history. Discussion of Sunni and Shi'a doctrine and thought in particular amounted to just 5% combined, and less than 1% discussed Sufism. In one cluster however, Islamic Discourse, theological discussions are prevalent (68.2%), and criticism of other faiths is significant (27.3%).

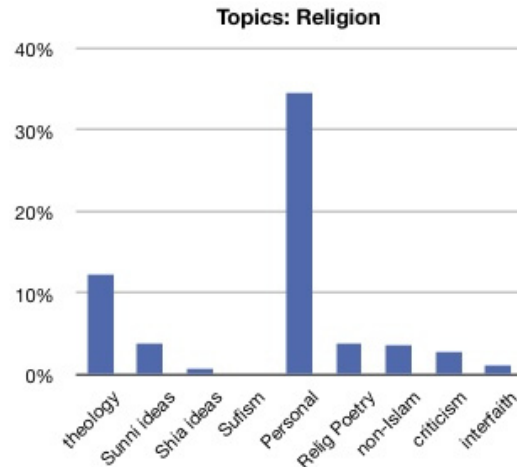


Figure 4: Religion

The majority of Arabic bloggers (73%) use blogs in a personal diary format (activities, private thoughts and reflections, etc), which is not surprising given the infrastructure of blogs. However, it is quite rare for them to discuss their personal single life (14%), which included love, relationships, dating, etc., or even family life (love, marriage, children, etc.) at just (15%).



Figure 5: Personal Life

Among cultural topics, 'high culture topics' (51%), including poetry, literature, and art, handily beat out pop culture topics, including movies and television (12%) and music (7%). Human rights (civil and political rights)²⁹ is a fairly popular topic, discussed among 43% of bloggers. Criticism of Western culture and values was fairly rare, discussed just 9% of the time, although support was equally limited (10%).

²⁹ By human rights we mean civil and political rights (freedom of speech, assembly, etc.). We did not indicate this on the code sheet, however, so it is possible coders could have interpreted the term more broadly.

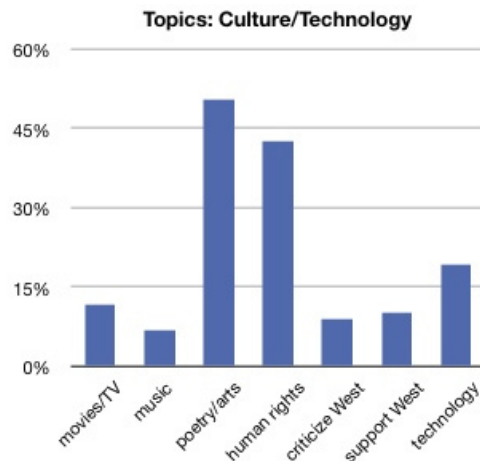


Figure 6: Culture and Technology

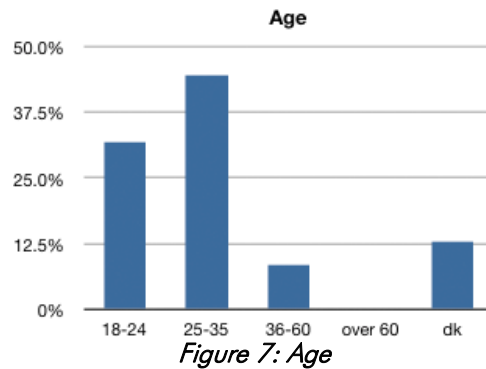
Demographic Results

A clear majority of bloggers (83%) are located primarily in their own country. Of the expatriates (5% total), just 3% are located outside of the Arabic speaking world. Coders could not tell where bloggers were located 13% of the time, which is not insignificant, but still much lower than in our Iran study. Specific country locations and concentrations of expatriates are discussed in the cluster descriptions, but we note here that Egypt has the largest number of bloggers on our map.

Since the Arab speaking world encompasses such a large, varied group of countries it is dangerous to generalize about freedom of speech in the Arab world, but we note that, as a region, the Middle East is one of the least free in the world according to global democracy and governance indices such as those from the World Bank, UN Arab Development Report, Freedom House, and the Economist Intelligence Unit. However, despite the low levels of freedom of speech in the region, 64% of bloggers use their name when writing, while only 37% blog anonymously, or use an obvious pseudonym.³⁰

In terms of age and gender, bloggers are overwhelmingly young and male. About three-quarters of bloggers (Figure 7) are under the age of 35, with the majority of those (45%) in the 25 to 35 year-old age range, while just 9% are over 35, and almost none are over 60 years of age, indicating that a significant gap between younger and older bloggers. Thirteen percent of the time coders could not assess a blogger's age range. Sixty percent of bloggers are male and just 34% female, indicating a significant gender divide, although this is still a higher proportion of female bloggers than we found in our Iran study.

³⁰ According to Marc Lynch anonymity has been a divisive issue in the Arabic blogosphere even though it can be risky for political bloggers to write with their names. "Many of the more politically engaged bloggers have chosen not to use pseudonyms, on the assumption that they would be unlikely to remain unknown to state security services anyway." Marc Lynch, "Blogging the New Arab Public."



Gender and Blogs

We broke out differences between known male and female bloggers to see if there were significant differences among blogger demographics or topical preferences. One of the most striking differences is found in anonymity results. Coding data reveals that while both groups were more likely than not to attach their name to their writing, female bloggers are much more likely to blog anonymously than their male counterparts. Forty-two percent of female bloggers wrote anonymously or with an obvious pseudonym, compared to just 29% of males.

While both male and female bloggers are young (under 35), female bloggers made up a much larger proportion of 18-24 year olds. The plurality of female bloggers (47%), were in the 18-24 year old category, compared to just 24% of male bloggers. There were also far fewer female bloggers over 35 (only 3%), compared to 14% of male bloggers. The majority of male bloggers were in the 25-35 year old category.

Among politics and public life issues, both men and women preferred to talk about domestic and international news more than anything else, in almost equal number. However, women's issues (rights, status, hijab, etc.) were talked about more often among female bloggers (16%), compared to males (6%).

Among personal life topics, females were more likely than males to discuss single life (19%) and family life (15%), but not by much; just 5% more than men for single life and 3% more than men for family life.

When talking about religion, female bloggers were more likely to discuss personal religious thoughts and experiences than males (61% vs. 53%), while men discussed the Qur'an, theology, and the interpretation of Islam slightly more often than women (21% vs. 18%).

Among all bloggers, a) poetry, literature, art, and b) human rights are the most popular topics in the culture and values category. However, female bloggers were much more likely to discuss poetry, literature, and art than men (47% vs. 30%), making it the most popular topic among women, while the second most popular for men. The most popular topic for men was human rights (31%), which was discussed less by female bloggers (23%). Male bloggers were also more likely to discuss technology (14%) than women (8%). Men were also slightly less likely to be critical of western culture and values (7% compared to 4%), but not by much, and this was not an especially popular topic overall for either gender group.

Female and male bloggers did not differ much in their preference for discussion of specific political topics. Male bloggers were slightly more likely to be critical of domestic political leaders than female bloggers (19% vs. 15%), and women were slightly more likely to voice support for Palestine than men (14% vs. 11%). Overall, however, their preferences among specific political topics were quite similar.

Gender differences are also discussed in the detailed cluster descriptions, but we note here that the Egyptian Youth sub-cluster (47% F) and Saudi Arabia (46% F) had the most female bloggers, while the Maghreb French (89% M) and Syrian (87% M) clusters have the largest proportion of male bloggers in the Arabic blogosphere.

THE ARABIC LANGUAGE MEDIA ECOSYSTEM

Like blog networks around the world, the Arabic blogosphere directs a lot of attention (via links) to traditional (mainstream print and broadcast) media sources.³¹ The patterns of these links reflect the divisions and clusters we see in the network, and also reveal a small set of sites that have broad reach across the Arab world. Many bloggers in the large Arabic language national clusters link to their own national media entities. For instance, mainly Egyptian bloggers link to www.ahram.org.eg, akhbarelyom.org.eg, and alwafd.org. Mainly Kuwaiti bloggers link to alwatan.com.kw, alqabas.com.kw, and alraialaam.com. And mainly Saudi bloggers link to alriyadh.com and alwatan.com.sa (Figure 8). These national media entities do receive some links from outside their co-national blogger cohort, but far fewer than from within. There is a very different story among the Levant/English Bridge bloggers, who overwhelmingly link to international English language media sources such as news.bbc.co.uk, guardian.co.uk, cnn.com, english.aljazeera.net, and nytimes.com (Figure 8). Some national, mainly Lebanese, media sites receive links from a significant number of bloggers as well, including the English language dailystar.com.lb, and Arabic language al-akhbar.com. The English language Israeli news site haaretz.com also receives links from a significant number of bloggers.

Looking at the sites that have broad reach across the Arabic blogosphere, the importance of ‘Web 2.0’ sites is striking. Of the ten sites with the most links from bloggers in our map (Figure 8), half are the same Web 2.0, user generated content sites popular in the United States and globally. YouTube (#1) is linked to by more than twice as many bloggers as the second most popular site, English language Wikipedia. English Wikipedia (#2) and Arabic Wikipedia (#6) each receive links from bloggers across the map (Figure 10). After Web 2.0, the second most important category of Web site is broadcast media. Al Jazeera (#3), BBC (#4), and Al Arabiya (#7) all reach broadly across the entire Arabic blogosphere, though BBC’s links are slightly weighted to the Levant/English Bridge cluster, Al Jazeera to Egypt, and Al Arabiya to Saudi Arabia. Perhaps it is worth noting that the US-sponsored Arabic

³¹ For an overview of traditional and new media in the Arab world and political impacts see Marc Lynch, *Voices of the New Arab Public: Iraq, Al-Jazeera, and Middle East Politics Today* (New York: Columbia University Press, 2006); William Rugh, *Arab Mass Media: Newspapers, Radio, and Television in Arab Politics* (Westport: Praeger, 2004); and Jon Alterman, “New Media, New Politics? From Satellite Television to the Internet in the Arab World,” The Washington Institute for Near East Policy, Policy Paper #48, 1998.

media outlets are far less successful among Arabic bloggers than regional efforts and the BBC, with Radio Sawa ranked at #507 and Al Hurra at #2,871. Two 'Web native' sites round out the top 10: Islamonline.net is a bi-lingual (English and Arabic) site featuring Islam-related international news and information, including applying Sharia in contemporary life, relations with other non-Muslim faiths, etc. Manalaa.net (#10) is a first-generation Egyptian blog with posts in English and Arabic, which by virtue of early-mover advantage is a staple of blog rolls across the Egyptian clusters, and among some outside as well. Of the top ten sites, it is the only one with links heavily skewed from a particular part of the map, in this case Egypt. Some additional sites are pan-Arabic in their reach, such as the news sites news.google.com (#21) and arabic.cnn.com (#23) and the Islam-oriented site Islamway.org (#15). Outside of the top ten, national or cluster-specific outlinks are the most common.

rank	#blogs	url	type
1	2070	youtube.com	Web 2.0
2	932	en.wikipedia.org	Web 2.0
3	817	Al Jazeera.net	broadcast
4	816	news.bbc.co.uk	broadcast
5	795	flickr.com	Web 2.0
6	667	ar.wikipedia.org	Web 2.0
7	566	alarabiya.net	broadcast
8	500	islamonline.net	webnative
9	483	digg.com	Web 2.0
10	456	manalaa.net	blog

Figure 8: Top 10 Widely Linked Sites

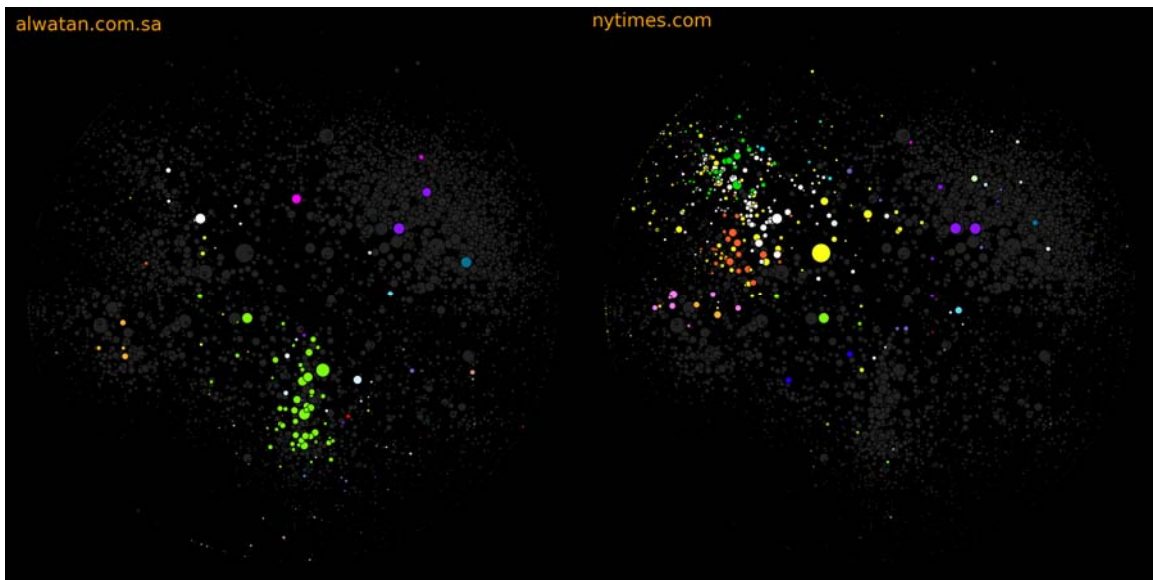


Figure 9: Links to Al Watan (Saudi) and New York Times

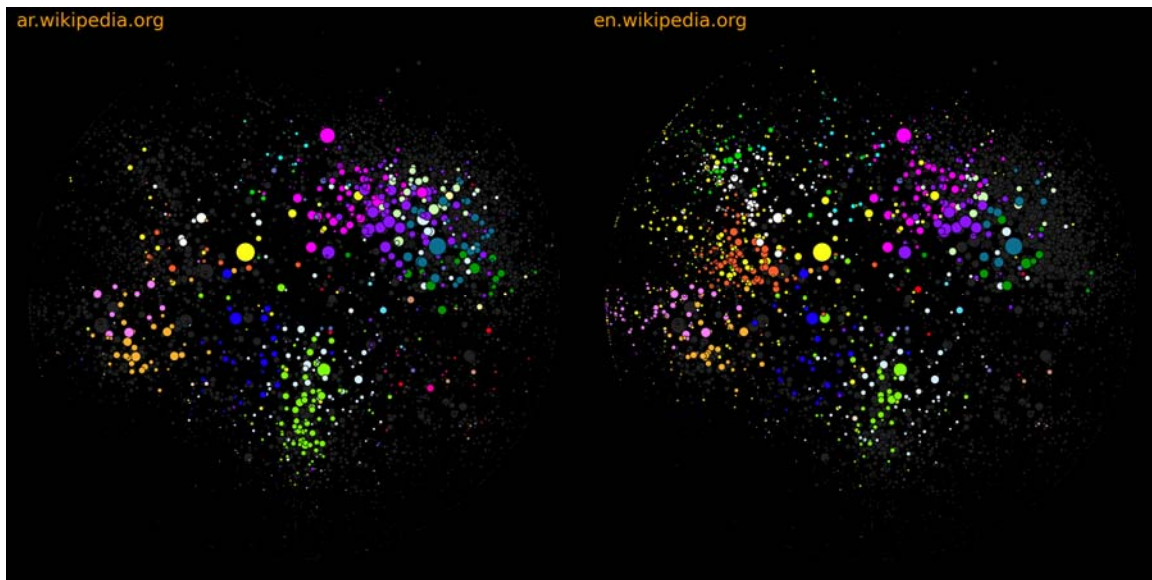


Figure 10: Links to Arabic and English Wikipedia

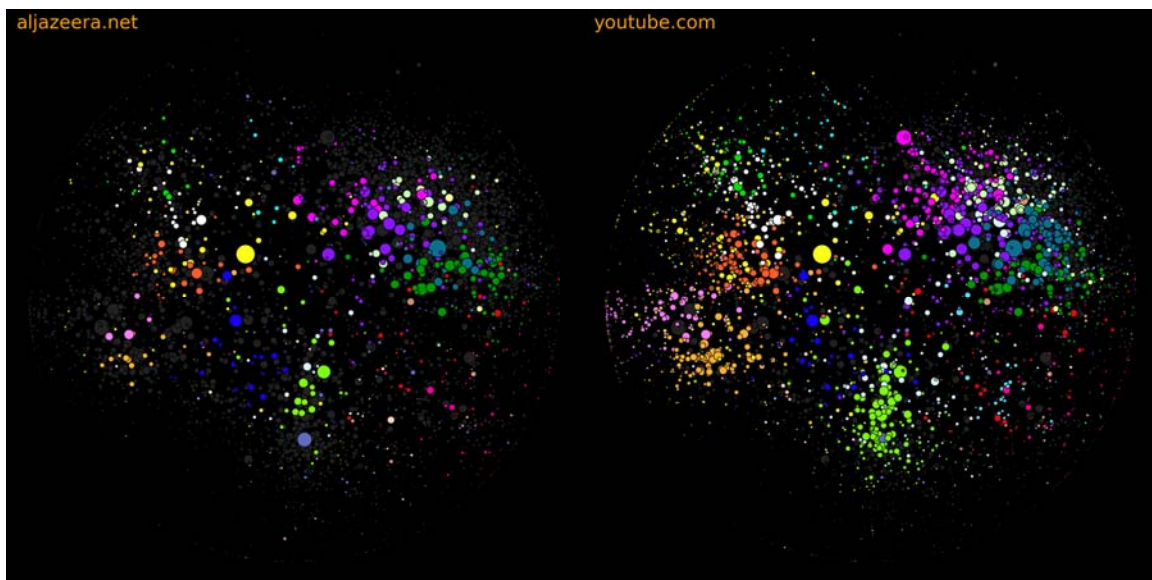


Figure 11: Links to Al Jazeera and YouTube

A CLOSER LOOK AT YOUTUBE

As noted above, YouTube is linked to by more Arabic bloggers than any other site, by a factor of more than two. However, there is a critical difference between YouTube and a news site like Al Jazeera that affects how it may or may not act as a central hearth for the Arabic networked public sphere. YouTube functions more as a media platform than as a publisher with editorial control, and so while Arabic bloggers link to YouTube videos in droves, they are not often linking to the same videos, and usually not in pan-Arab patterns. Instead, the distribution of links to particular videos mirrors in microcosm the distribution of links to Web sites. Very few videos are linked to by bloggers spread across the map, while

more are linked to by bloggers from particular clusters. We looked at 30 of the most highly cited recent videos and found that the content tends to be politically oriented. And there is one topic in particular that is the focus of pan-Arabic blogosphere interest: the Palestinian-Israeli conflict, particularly the conflict in Gaza. The most highly cited video, posted on January 07, 2009, features a song (in English, by an American) against the Israeli actions in Gaza, including video clips of bombings. Among the top 30, eight are about the Israel/Palestine conflict, and six are about Gaza in particular (however, one of these is a separate upload of the song that is #1.). The second most popular video, for which there are also two versions in the top 30, is the one in which a shoe is thrown at George W. Bush in Baghdad.

Aside from Palestine/Gaza and the projectile shoe incident, the remainder of the top 30 are mainly videos with cluster-specific political or social significance. These include clips recorded from television, produced videos, and amateur video shot on the street of unfolding events.

One shaky amateur video shows Saudi Shi'a minority women attempting to perform a ceremony in Medina, while groups of youth allegedly desecrate ('steal soil') from the graves of revered Islamic figures buried there. Dr. Aseel Al-Awadi, a professor at Kuwait University, announces her candidacy for the Kuwaiti parliament. There are a number of videos about Egyptian Baha'i, including incidents of discrimination and events affecting their rights. These are mainly linked to by the small cluster of Baha'i bloggers on the map, which demonstrates the ability of small focused groups to drive link counts, though not always view counts. These videos and others show people and events with mainly local or national implications, although they feature common themes such as the abuse of power, interpretation of Islam, and religious tolerance.

num	count	YT video ID		title
1	50	dlfhoU66s4Y	pan	We will not go down (Song for Gaza) -by Michael Heart
2	29	OM3Z_Kskl_U	pan	Raw Video: Iraqi Journalist Throws Shoe at Bush
3	20	3usdvcl1fVQ	eg	TV clip about controversial Baha'i new year celebration
4	19	4EWPZ5sm4t4	kw	I swear...! Kuwaiti Celebrity PSA campaign
5	17	ewXkEHNYyB8	pan	a trailer for a documentary about occupation of Gaza
6	12	h-E4-E63ok0	egy	'The Pasha's Daughter is Terrifying People on the Street'
7	12	d8NZTogDosc	kw	female Kuwaiti professor announces candidacy for Parliament
8	11	Vn-iT70e8bU	bah	Baha'i homes burning, incident from April 2009
9	11	o5UAkDjgev8	kw	'In Kuwait, Singing Is Still Possible' separation of sexes protest
10	11	Hqzw7oBZT8k	pan	Animation about restrictions of Palestinians' movement
11	11	VVZtcW6HOco	sa	Saudi Prince calls into sports TV show, harsh words for host
12	10	1LjgfUwlp4	eg	Egyptian TV report of Baha'i house burning incident
13	10	iszbLbUf-OA	pan	Young girl recites poem about Palestinian suffering
14	9	H39i5tYTxaI	pan	Al Jazeera clip of Palestinian child describing Gaza attack
15	9	zGp2we8bVac	kw	Kuwait Parliament member at Hamas celebration, criticized
16	8	jAz6jPg9LiM	sa	Al Jazeera discussion of Internet use for 'resistance' in Gaza
17	8	rpiccERJaFk	sa	Video reply to Geert Wilder's anti-Islamic film, 'Fitna'
18	8	DeaZuj7ruwM	pan	Fox News interview with Rabbi from Jews Against Zionism
19	8	BmgfQiaYEdk	pan	Saudi Cleric denounces mice
20	8	M8GOrc0-Ygg	pan	another video of Bush shoe incident

21	8	IHS0P7aZZ4o		video removed for copyright infringement
22	8	Jbr8LBR5k-8	egy	repeat of video #1, 'We Will Not Go Down'
23	8	AlzbwV7on6Q		Trailer for Slumdog Millionaire
24	8	0xOLy9k_6VA		Al Jazeera program on Arabic (mainly Levantine) blogosphere
25	8	JXhpFSb8mQ0	egy	Mostly silent clip of footage inside of Egyptian prison
26	8	TCj7sY9Tn9A		Al Jazeera clip on Turkish President's frustration at Davos
27	8	AnwQ_PhDr9c	lb	Lebanese TV presenter screams at Hezbollah and Leb. Army
28	7	ObctpzziUGo	eg	Egyptian TV announcing Court decision on Baha'i rights
29	7	39gAH4pYOSI		'Island Talk' Former prisoner questions Guantanamo
30	7	S_BY39tnOkw	eg	Saudi Shi'a in Medina, incident captured on phone cam

Figure 12: Top 30 Most Cited YouTube Videos

Examples of Videos linked from across the Arabic blogosphere

1. <http://www.youtube.com/watch?v=dlfhoU66s4Y>

Title: We will not go down (Song for Gaza) - by Michael Heart

Summary: This video clip is of a song (in English) against the Israeli actions in Gaza, including video of bombings. Some of the lyrics are as follows:

*A blinding flash of white light
Lit up the sky over Gaza tonight
People running for cover
Not knowing whether they're dead or alive*

*They came with their tanks and their planes
With ravaging fiery flames
And nothing remains
Just a voice rising up in the smoky haze*

*We will not go down
In the night, without a fight
You can burn up our mosques and our homes and our schools
But our spirit will never die
We will not go down
In Gaza tonight*

2. http://www.youtube.com/watch?v=OM3Z_Kskl_U

Title: Raw Video: Iraqi Journalist Throws Shoe at Bush

Summary: The second most popular clip in the Arabic blogosphere is Associated Press coverage of Iraqi journalist Muntadhar al-Zeidi famously throwing both his shoes at then U.S. President George Bush at a press conference in Baghdad.

5. <http://www.youtube.com/watch?v=ewXkEHNYyB8>

Title: Why the Children of Gaza Do Not Deserve to be Killed?

Summary: This is a trailer for a documentary called 'Occupation 101' about Israel's occupation of Gaza, as told through the eyes of Palestinian children.

10. <http://www.youtube.com/watch?v=Hzqw7oBZT8k>:

Title: Closed Zone.

Summary: This animated clip was created by Gisha (The Legal Center for Freedom of Movement), an Israeli organization. It shows the limitation on the movement of Gazans due to Israeli security restrictions. According to the clip's description on YouTube, "Gisha - Legal Center for Freedom of Movement calls on the State of Israel to fully open Gaza's crossings and to allow the real victims of the closure - 1.5 million human beings - the freedom of movement necessary to realize their dreams and aspirations."

14. <http://www.youtube.com/watch?v=H39i5tYTxaI>

Title: A Palestinian child describes the children's suffering during the war on Gaza. (From Al-Jazeera Arabic TV).

Summary: The child in this clip describes how the children of Gaza have suffered from the attack on Gaza. According to experts in the region, this clip was very popular because the child was very articulate. He talked about how the children have no water and food, and that they watch bloody conflicts on TV instead of cartoons and children's programs. Al-Jazeera TV used part of the clip as a promo during the war on Gaza.

18. <http://www.youtube.com/watch?v=DeaZuj7ruwM> (In English)

Title: Interview with Rabbi Yisroel Weiss

Summary: This is a clip of Fox News commentator Neil Cavuto interviewing Rabbi Yisroel Weiss from Jews United Against Zionism about whether it is proper for Jews to have their own political state.

Examples of Videos linked by particular clusters:

Kuwait

4. <http://www.youtube.com/watch?v=4EWPZ5sm4t4>

Title: I swear...!

Summary: This clip comes in the context of a campaign called 'I swear', launched by a Kuwaiti media production company called *Zawaya*. 'I swear' was launched this year (2009) on February 25, which marks national Kuwaiti day. The clip consists of Kuwaiti celebrities pledging to do good things for the sake of their country. According to experts in the region, this video was initially welcomed due to its patriotic nature, and it was also found to be very similar to a one produced by Barack Obama during the US presidential campaign (<http://www.youtube.com/watch?v=4EWPZ5sm4t4>)

7. <http://www.youtube.com/watch?v=d8NZTogDosc>

Title: Dr. Aseel Al-Awadi Announcing Her Candidacy for the Kuwaiti Parliamentary Elections

Summary: This is a video of Aseel Al-Awadi, a professor at Kuwait University, announcing her candidacy in the May 2009 Kuwaiti parliamentary elections. In the video she argues for better health care for Kuwaitis while also criticizing the parliament for focusing on 'less important' political issues, while neglecting critical causes.

9. <http://www.youtube.com/watch?v=o5UAkDjgev8>

Title: In Kuwait... Singing Is Still Possible

Summary: The video is against legislation passed in 1996 that banned the mixing of sexes in classes, libraries, cafeterias and other public places. It includes a popular song, 'Singing Is Still Possible,' which was performed in the movie *The Destiny*. It includes press headlines such as 'YouTube is Banned in Kuwait' and 'Minister Calls for Censoring of the Internet' as well as clips of women using computers, sitting shoulder to shoulder next to men during a seminar, and a march by women in protest against the law while holding banners which announce "we will not allow fanatics to control us" and "Islam is clear... Sharia preserves our freedoms." The clip also shows news headlines on the broad censorship of books in Kuwait, including one that wonders, "Are books considered weapons of mass destruction?"

15. <http://www.youtube.com/watch?v=zGp2we8bVac>

This clip has been removed by the user. Here it is in an alternative location:

<http://www.youtube.com/watch?v=VqaZmL1jIxY>

Title: Hamas celebration in Kuwait

Summary: This video shows a victory celebration organized by Hamas supporters in Kuwait. A Kuwait Islamist Parliament Member, Naser al-Sanee, was allegedly among the attendees. According to experts from the region, this upset many Kuwaitis, some of who expressed anger in their blogs, because the bloggers said Hamas supported Saddam Hussein's invasion of Kuwait.

Egypt

6. <http://www.youtube.com/watch?v=h-E4-E63ok0>

Title: The Pasha's Daughter is Terrifying People on the Street

Summary: The video shows a woman fighting with a man, in a street at Giza neighborhood in Egypt. She explains to the people who were watching the fight that she offered compensation to the man, that she hit his car by mistake, but that he started to insult her.

The woman is seen calling someone, who, according to her, is her father who is a the one high rank officer in a Giza security body. She is heard asking him to send her one of his police officers. Also, she is holding a stick on her hand, and someone in the clip says, "She told me that this is an electrical stick, and a beat with this stick would paralyze me!"

The video is titled 'The Pasha's Daughter...' because in the Egyptian dialect 'Pasha' refers to a police officer.

According to experts from the region, this video was first published in last August (2009) by the prominent blogger Wael Abbas after it had been filmed by someone on the street with his video camera. The blogger was able to obtain the license plate number of her car from the video and called for investigation into the woman. An independent newspaper (Al-Masry Al-Youm), managed to find the man who appears in the clip, whose car was struck, and who said that he followed her followed her car to report the accident to the police, and her mother eventually apologized on the woman's behalf and he accepted her apology. Many in Egypt were upset that the women was abusing her father's position in power, as well as the language she used against the person she had hit with her car.

25. <http://www.youtube.com/watch?v=JXhpFSb8mQ0>

Title: Prisons in Egypt

Summary: Mostly silent clip of inside of Egyptian prison; the speaking is very difficult to understand, partly due to dialect and partly due to sound quality.

Saudi Arabia

11. <http://www.youtube.com/watch?v=VVZtcW6HOco>

Title: Prince Sultan Bin Fahd ... important words

Summary: This video is a clip from the Saudi Sports channel show dedicated to analysis of the defeat of the Saudi national football (soccer) team's loss to Oman during the Gulf championship. After discussing the loss, Prince Sultan Bin Fahd, the head of Saudi league for football joins the conversation by phone and quickly upbraids the panel one by one for their criticism of the team, including, "You! All what you said is wrong", "Don't speak about something you don't know... you're ignorant!", "If you give valuable advice, we will listen." The Prince then singles out the most critical panelist saying "Behave yourself!...Talk about something you understand, or keep silent", "We let you evaluate and analyze, that does not mean you go on criticizing out national team", "We lost the score but won good players!", and "behave like Saudis."

The silenced panel meekly replied with comments like "God bless your lifetime!" and only one dared to comment on the prince's verbal assault by saying, "I am criticizing technical points about the team," which lead to another outburst by the Prince. "Behave yourself! If you were ill bred I would teach you what good manners are," but the man replied again "I am well bred, and not criticizing the kingdom, but a national sport team, which you are responsible for." The Prince got angry again, says unclear words before hanging up, the host thanks the Prince, and the clip ends.

30. http://www.youtube.com/watch?v=S_BY39tnOkw

Title: Untitled

Summary: This clip shows a group of women from the Saudi Shi'a minority group performing some rituals at a cemetery in the city of Medina, the second holiest Muslim city, where revered Muslim religious figures are buried. According to the Saudi authorities, the Shi'a women were performing rituals offensive to Muslims and tried to prevent them from doing so. Clashes between Shi'a groups and Saudi religious police erupted because of this incident. Shi'as claimed that the riot police were heavy-handed in dealing with the women and the other Shi'a protesters. In the foreground, youth allegedly 'stealing soil' from graves.

Baha'i

8. <http://www.youtube.com/watch?v=Vn-iT70e8bU>

Title: Baha'is (Baha'is Houses Burning)

Summary: This short video shows five Baha'i houses that were set on fire in April 2009, in a small island called Shuraneyah in Sohag governate in Upper Egypt. In this video, nothing but the fires and some shadows of people surrounding the burning houses is shown. Muffled voices are heard saying, "There is no God but Allah" and in the end that the police are coming.

12: <http://www.youtube.com/watch?v=1LjgfUwilp4>

Title: Burning the houses of the peaceful Baha'is. (A clip from Egypt's al-Mehwar TV.)

Summary: This is a clip from Egypt's al-Mehwar TV about attacks on the Baha'is. The show discusses who is behind attacks and the burning of Baha'i houses just one day before a conference that was organized by the Baha'is.

DISCUSSION AND CONCLUSION: DIVERSITY AND DISCOURSE IN THE PUBLIC SPHERE

In a part of the world where print and broadcast media traditionally have been controlled by the government, digital networked spaces offer the possibility of a much richer public sphere than existed before, providing a venue for all manner of opinions and ideas, covering love, life, culture, economics, religion, and politics. The methodology and findings of this study are grounded in the quintessentially public nature of the blogosphere. Unlike many other sequestered neighborhoods of online and offline life, there are relatively few institutional obstacles to becoming part of the audience or joining the conversation. The blogosphere explicitly welcomes readers. Blogs neither charge for admission nor require readers to pass through a vetting process for admission to the club. Blogs tend to offer archived, permalinked units of content that are visible to anyone, designed to link and be linked to, and tagged with metadata for maximal visibility to search engines. This participation in the link economy makes blogs an important part of an ongoing public record, indexed by search engines, cited by other bloggers, and otherwise kept alive as engrams in the social mind. The arguments, opinions, and analyses that are found in the blogosphere have been intentionally offered up to the world, an open contribution to public dialogue.

There are many areas of the Internet that are less public. The more limited, ephemeral modes of communication found in forums, chat rooms, bulletin boards, and social network services do not offer the same searchable, more permanent public record but rather represent different varieties of private arenas, often in the shadows of the networked public sphere. In these we are more likely to find personal social conversations, as well as the subversive and the profane, not significantly linked to the larger public discourse, sometimes intentionally lurking on the fringes, and always less visible to the outside. The communication in these quasi-private realms is limited in audience and generally excluded from the 'link economy' that knits together much of the content on the Internet. Generally speaking, these enclaves are well suited for low-profile discussion among networks of friends and like-minded strangers, and are sometimes an organizing space for mobilization of these networks. The architecture and social norms of the blogosphere, on the other hand, are conducive to public expression, contestation, and synthesis of ideas. Blogging is not the main way that Internet users in the Arab world interact online; Arabic forums are extremely popular, as are Facebook and other 'hybrid' platforms, like Live Spaces, which combine online journals with platform-specific social networking functions and do not (in practice)

connect strongly to the wider online public. However, where it thrives, blogging creates open, society-wide networks that generate, filter, and circulate information, feeding back into the mainstream media and directly to political elites.

The topics and ideas that disseminate in the blogosphere are the result of choice and interest in a menu of issues that have been intentionally offered for public consumption. Seen through this lens, the clusters that emerge in the Arabic blogosphere and the issues that occupy this public space are indicative of large social processes and preferences, and are highly suggestive of the interests and mood of the public. Further research is required to determine how opinions in blogs correlate with offline public opinion, and the degree to which these online publics may represent 'counter-publics' speaking online because they have little outlet offline.

The formation of nation-based clusters, where networks of bloggers connect with fellow citizens, is a key finding of this research. The logic of this structure mirrors the reality of political and economic life, being rooted primarily in national and local affairs. The relative paucity of pan-Arab dialogue in the blogosphere, outside of where Arabic bloggers engage the West, is also notable. Our research shows that bloggers in the Arab world focus more attention on local political issues and leaders, and when discussing their leaders, devote more energy to criticism than support, a critical point of departure from the mainstream press. The conventional wisdom has been that Arab leaders have sought to deflect criticism of their own regimes toward international issues, including those promoted by extremist groups. The results of this study suggest that the opposite happens in the Arabic blogosphere, where more developed national networks focus more attention inward.

Those online structures that have not emerged are as telling as those that do. Although political ideologies play a substantial part in the formation of online communities, the clusters that we observe do not correspond to political parties. The emergence of online political clusters without clearly defined offline counterparts points to both the power of this new platform and its limits. As long as they are restricted to the online world, these emergent organizations will be constrained in the power and influence that they can exert.

The political movements of Egypt, such as Kefaya and the Muslim Brotherhood, are possible exceptions to this trend, as with them we see evidence of online groups coalescing around distinct political and religious ideas. This may be a sign of the Egyptian blogosphere's complexity, and we would expect to see this elsewhere as Arabic online publics grow. The case of the Muslim Brotherhood is especially interesting. While they are not permitted to openly organize offline, many in the Muslim Brotherhood have taken their ideas to the blogosphere. Opening up their thoughts to criticism within and outside their own community suggests a degree of confidence in the resilience of their ideas and amenability to debate. It is interesting to see these decidedly democratic practices taking root in civic organizations online.

Although not as important as domestic issues, international political topics are well covered in the Arabic blogosphere with very similar chords being struck across the region. Support for the Palestinian cause appears nearly universal, as is criticism for Israel. US involvement in the region is largely seen in negative terms. However, the relatively modest level of attention paid to the United States might come as a surprise to many Americans.

In press and policy circles alike, a lot of attention has been paid to the potential of the Internet for terrorist recruitment and propaganda, and a lot of this concern is focused on the Arab world. However, our study indicates that Arabic language blogs are not to any significant degree used to support extremism, preach hate, or organize terrorist activities.³² In fact, support for terrorism or violent confrontation with the West appears to be rare. Out of the over 4,000 Arabic blogs that we coded, only a handful were judged to support violent extremism.³³ Furthermore, criticism of terrorism is common across the Arabic blogosphere. Some research has shown that even in Arabic language forums, which are less public than blogs in important ways and also provide more anonymity, terror discussions are a minority activity.³⁴ According to Islamic Internet scholar Gary Bunt, “Participation in militaristic jihad is a minority issue, on- and offline. Muslim individuals and organizations have expended considerable energies—on the Internet and elsewhere—distancing themselves from such acts.”³⁵

We do not argue that extremist Web sites and forums do not exist; certainly, they do, and our research does not address their impact. However, academic studies and media reports that focus exclusively on terrorist uses of the Web can leave the impression that this is a dominant form of discourse in the Arabic language Internet, and could lead to ill-informed policy responses, which could unintentionally limit the diverse, open, and often civically-minded political, cultural, and religious discussions that take place in blogs and other open Internet spaces.

³² For a discussion of how terrorists and Jihadists use the Internet see Bruce Hoffman, *Inside Terrorism* (New York: Columbia University Press, 2006) 197-228; Gary Bunt, *iMuslims: Rewiring the House of Islam* (Chapel Hill: University of North Carolina Press, 2009) 177-241; Yariv Tsfati and Gabriel Wiemann, “www.terrorism.com: Terror on the Internet,” *Studies in Conflict & Terrorism*, 25 (2002); and Evan Kohlmann, “The Real Online Terrorist Threat,” *Foreign Affairs*, Vol. 85, Issue 5, (September/October 2006) available at <http://www.foreignaffairs.com/articles/61924/evan-f-kohlmann/the-real-online-terrorist-threat> (accessed May 27, 2009).

³³ The definition of ‘terrorism’ is of course problematic, and beyond the scope of our research. Clearly, interpretations of what constitutes ‘terrorism’ vary across the Arab world, and the clash of interpretive regimes shows up in blogs, which can criticize terrorists in one post, and praise Hamas’ ‘resistance’ in another. In this paper, we simply make a distinction between support for anti-Israeli political organizations like Hamas and Hezbollah that are considered terrorist organizations by many countries and international bodies, but are often seen differently in the Arab world, and support for organizations like Al Qaeda and violent extremist philosophies. The latter is what we refer to here as “terrorism,” leaving the former as an issue for discussion by more knowledgeable specialists.

³⁴ Helmi Noman found that the majority of the forums he studied in 2005 were Islamic-themed. Regarding Jihadi forums, he writes, “These are few in numbers, but extremely active. They are believed to be run by Jihadi organizations and their supporters. Content in these forums is mostly claims of responsibilities and Fatwas on Jihad. Media organizations and several Western research sites actively monitor these sites and report their newsworthy materials. Authorities usually shut down these forums but they continue to pop up in new web servers.” See Helmi Noman, “Content and Usage of Arabic Online Forums and Groups,” available at: http://www.helmionline.com/Internet/2005/10/web_content.html#more (accessed May 20, 2009).

³⁵ Gary Bunt, *iMuslims: Rewiring the House of Islam* (Chapel Hill: University of North Carolina Press, 2009) 183. See also Hsinchun Chen, Sven Thoms, T.J. Fu, “Cyber Etremism in Web 2.0: An Exploratory Study of International Jihadist Groups,” IEEE International Conference On Intelligence and Security Informatics, 2008 (Forthcoming) available at <http://ai.arizona.edu/research/terror/publications/ISI2008-Sven-WEB2.pdf> (accessed May 27, 2009), who found only 28 Jihadist blogs in an exploratory study using automated techniques; most of those blogs were in Arabic and located in the Middle East.

When allowed to flourish, blogs can contribute to the formation of a networked public sphere with significant political, social, and economic value. As articulated by Yochai Benkler, the networked public sphere can support efficient interaction among important communities of interest, allowing them to negotiate agendas, process opinion, and broker knowledge. While these may sound like abstract values, blogging already shows signs of having an impact in the Arab world, such as facilitating cooperation between secular and Islamist reformers in Egypt, and promoting discussion and reform within the Muslim Brotherhood.

A popular meme is the notion that digital tools will play a key role in the reform of authoritarian political systems and will help to promote democratization. At the core of these hopes are several universally compelling ideas that are clearly at work in the Arabic blogosphere: opening up avenues for greater political participation, facilitating transparency in governance, empowering individual voices and minority viewpoints, and democratizing the production and dissemination of information.

As promising as these possibilities are, the emergence of the Internet and digital technologies does not in itself constitute a one-way road towards political liberalization. The ideas and debates of the blogosphere are not effectively open to all. Many have no access to the Internet, and linguistic obstacles prevent most from easily tapping into the full range of ideas available online. Furthermore, effective engagement in the field of ideas requires technical skills and media savvy, which are not uniform even among those with regular access to the Internet. Moreover, governments have acted to limit the influence of this digitally-mediated public sphere. Numerous countries in the Arab world have turned to filtering political speech on the Internet that they deem objectionable. In Egypt, where there is no filtering, bloggers have been arrested and imprisoned for their writing.³⁶ Government authorities continue to search for mechanisms to control cyberspace while activists and dissidents develop new ways to avoid these controls.

Even where digital tools might be used to wrestle power from the hands of authoritarian governments, it would be a mistake to assume this would naturally produce classically Western versions of liberal democracies based on individual liberties, freedom of belief and expression, and a secular political order. Hundreds of Egyptian bloggers forcefully advocate for human rights, freedom of speech, and an end to corruption and torture. But two of the strongest groups among them are radical leftists, the West's great enemy of the last century, and Islamists, feared by some as the great threat of the new century.

While we caution against simplistic versions of what could be called the 'techno-democracy' argument, there are more sophisticated formulations that deserve further research. The Muslim Brotherhood that mobilizes mindshare in the networked public sphere is no longer the same Muslim Brotherhood. As we see with advocacy organizations in the United States, or Shi'a religious students in Iran, the move to Internet modes of communication can alter the forms of organization among people committed to similar goals, ideas, and values. The

³⁶ Ronald Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, *Access Denied: The Practice and Policy of Global Internet Filtering* (Cambridge: MIT Press, 2008) 276-279.

Internet does not just promise (or threaten) to change the balance of power among players on the field, it changes the field and changes the players too.³⁷

This study comes at an interesting juncture in US relations with the Arab World, in the wake of President Obama's call for a "new beginning between the United States and Muslims around the world." At the heart of this new beginning is a plea for understanding between cultures, to "listen to each other; learn from each other," to dismantle gross stereotypes, and to prevent extremist views from setting the tone of cross-cultural relationships. There is ground for optimism towards making progress on these accounts in the Arabic blogosphere. This research indicates that extremist views occupy a very small proportion of the ideas and opinions found there. Instead it is a space populated with a broad diversity of views, many of which promote common international values such as free speech and human rights.

³⁷ See Clay Shirky, *Here Comes Everybody: The Power of Organizing Without Organizations* (New York: Penguin Press, 2008).

APPENDIX A: NOTES ON METHODOLOGY

We used computational social network mapping in combination with human and computer content analysis to analyze the Arabic language blogosphere.

Baseline Data and Mapping

The basis of the social network analysis and blogs selection was a corpus of Arabic, French and English-language blog data collected since March 2008. We identified and collected data for over 35,000 Arab language blogs, excluding numerous user accounts for social network service ‘hybrids’ (which combine blog-like journal features with social network function like platform-specific friends lists and group memberships), like Windows Live Spaces and MySpace. The latter were found to form their own networks, mostly self-contained, linking very infrequently to blogs outside the particular service. Where journals were found on these services that did link to the wider blogosphere, they were tracked and remained part of the base collection used for the study. Initially, the study envisioned mapping only Arabic language blogs, but it was quickly discovered that a large proportion of Arab bloggers write in English, and mixtures of English and Arabic, with no clear network division between the two languages (unlike English and Farsi, for instance, between which there is a clear division). A sizeable number of blogs also mix French and Arabic. Therefore, lists of hundreds of known English and French Arab blogs, from expert sources and a range of Arab-oriented content aggregators, were used to seed network-based ‘snowball’ techniques to identify additional Arab bloggers writing in these languages as well. Because of the difficulty of separating Arab from non-Arab blogs in these ‘bridge’ regions, there are a small number of blogs authored by individuals without geographic, ethnic or family ties to the region that are included in the network because they are heavily interconnected with the Arabic language blogosphere. Social network analysis was used to identify the most connected blogs, the top 6,451 of which were mapped to identify the core structures of the Arabic blogosphere, create visualizations, and identify blogs for human and computational content analysis. The map of the Arabic blogosphere is plotted using the Fruchterman-Rheingold algorithm, which employs a ‘physics model’ approach in which blogs that are more densely connected are drawn together into clustered ‘network neighborhoods.’ The color of the blogs results from ‘Attentive Cluster Analysis,’ in which the linking histories of blogs are compared statistically in order to identify groups sharing similar preferences for linking to online resources.

Content Analysis

Once the network map was established, we next employed a mix of computer text and metadata analysis with human content analysis to determine basic characteristics of bloggers and their preferred topics across the network. The goal was to produce baseline estimates and findings that might inform more rigorous, targeted follow-on research.

- Computer text analysis: Using several types of lists, we searched full text of blog posts and compared the relative frequency of use of various words and terms across the clusters of the blog map. The three types of lists were: a.) all individual words in the text; b.) a list of over 200 terms, names and phrases chosen by expert advisers to represent key ideas and political actors; and c.) a list of over 53,000 Arabic language terms extracted from Arabic language Wikipedia, for which English language articles also exist. Aggregate cross-network use of key words/text (for example, *Israel*, *America*,

Gaza, Al Qaeda) was considered a rough indication of overall levels of discussion of key issues. Patterns in the use of these terms by various clusters were used to infer topical preferences of *attentive clusters*, with high-frequency terms considered indications of a particular cluster's interests and chosen topics relative to other clusters.

- Human blog coding: A sample of over 3300 blog URLs from across the map, as well as a group of blogs not included in the map (used for comparison), were coded from April 2008 to March 2009 using an online tool created by the Berkman Center. Data for all 4,370 blogs coded are included in the report, since differences between the results for the two data sets were minimal (differing at most 1%-3%, if at all). Ten Arabic-speaking researchers read blogs and coded them according to two standard scripts for blogs (attached as Appendix B). Outlinks, sources linked to by bloggers (usually Web sites, but also highly-linked blogs), were also coded with a separate script which is also included in Appendix B. Coders came from a range of countries in the region including Iraq, Egypt, Palestine, Kuwait, and Saudi Arabia, as well as Arabic speakers from Pakistan and the United States. The coding was initially intended to collect basic demographic data and topic data to supplement the computer text analysis, as well as to provide a richer qualitative sense of the discourse in Arab blogs. The latter was accomplished in part through the collection of text descriptions of the blogs, in which the research assistants were asked to describe blogs in one or several sentences. These descriptions were very informative. The collection of categorical data presented a number of challenges that suggest issues for methodological improvement in follow-on research. First, blogs were coded 'live,' i.e. the coder read the content that was on the front page of the blog at the time they coded it. Because of this, topics could fluctuate across the network over the time period of the study, which introduces noise into the results. Given that the posts read by researchers for a given blog might have differed, this variation impacts the inter-coder reliability measures for the data collected. Second, because of differences in dialects, slang and national aspects of culture and politics, not all coders could be equally sensitive to important aspects of speech in blogs in different parts of the network. This suggests the need for a combination of more specific questions and coders with special knowledge of key countries. While the latter presents its own problems for coding validity, we have found that the most valuable qualitative insights come from those coders attuned to the subtleties of the discourse they are monitoring. Given the repressive nature of some of the societies in question, the role of 'dog whistle' politics is likely to be strong. Third, some of the topics we cover, such as attitudes toward extremist violence, involve concepts with very different interpretations across the Arab world. Many of these regional variations are captured in the qualitative descriptions, but there are challenges to framing categorical attitude questions very specifically. For these reasons, we avoid drawing any conclusions in this study based solely on the data collected through human coding but instead draw on multiple sources of information, including the human coding, text analysis and network analysis, which is constructed on linking patterns of the blogs in the network.

APPENDIX B: CODE SHEETS

Arabic Extended

Is this URL a blog? (If not a blog, click no and jump to complete item at the bottom of the page)

- ☐ yes
☐ no

What is the nationality of the blogger?

Is the blogger located in his/her country, or an expatriate?

- ☐ in his/her own country
☐ expatriate (outside the Arab world)
☐ expatriate (in another Arabic-speaking country)
☐ don't know

Where are they located specifically (country, city, town, region)? If you list a city or town, you must also list the country. Please write 'dk' (for 'don't know') if you can't tell anything about where they are.

Does the blogger reveal their name, or write anonymously?

- ☐ uses name
☐ anonymous (including pseudonyms)

Is the blogger male or female?

- ☐ Don't Know
☐ Female
☐ Male

Can you estimate the blogger's age? (select the most appropriate category)

- ☐ 18-24 (student age)
- ☐ 25-35 (young adult)
- ☐ 36-60 (adult)
- ☐ over 60 (older adult)
- ☐ don't know

Please read the blog first, and then go back a second time for coding. The goal is to check off all topics that the blogger discusses, in the categories below. If the blog has a lot of posts, concentrate your reading and coding on the five to seven most recent posts. Check all topics that are discussed.

General Topics: Politics and Public Life

- ☐ International news
- ☐ Domestic news
- ☐ International policy
- ☐ Domestic political reform (advocating change within the current regime/system)
- ☐ Economic issues (gas prices, inflation, oil revenues, etc)
- ☐ Women's issues (rights, status, hijab, feminism, etc)
- ☐ Ethnic minority issues (rights, status, language, regional issues)

General Topics: Religion

- ☐ Islam (the Quran, theology, interpretation, history, etc)
- ☐ Sunni Islam (doctrine, thought, practice, leaders, etc.)
- ☐ Shiite Islam (doctrine, thought, practice, leaders, etc.)
- ☐ Sufism (doctrine, thought, practice, leaders, etc.)
- ☐ Personal religious thoughts and experiences
- ☐ Religious poetry
- ☐ Religions other than Islam
- ☐ Criticism of other faiths
- ☐ Interfaith dialogue

General Topics: Personal

- ☐ Personal diary (activities, private thoughts and reflections, etc)
- ☐ Single life (love, relationships, dating, etc)
- ☐ Family life (love, marriage, children, etc)

General Topics: Culture, History and Technology

- ☐ Movies and television
- ☐ Music
- ☐ Poetry, literature, and art
- ☐ Human Rights
- ☐ Western culture and values (criticism)
- ☐ Western culture and values (support)
- ☐ Technology (computers, software, blogging, etc.)

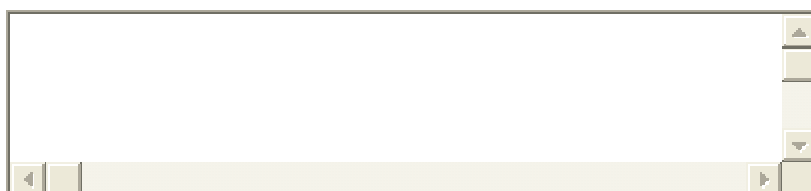
Specific Topics: Politics and International Affairs

- ☐ Domestic Political Leaders (Support)
- ☐ Domestic Political Leaders (Criticism)
- ☐ Foreign Political Leaders (support)
- ☐ Foreign Political Leaders (criticism)
- ☐ War in Iraq
- ☐ War in Afghanistan
- ☐ United States (Support)
- ☐ United States (Criticism)
- ☐ Palestine (Support)
- ☐ Palestine (Criticism)
- ☐ Israel (Support)
- ☐ Israel (Criticism)
- ☐ Terrorism (support)

- ☐ Terrorism (criticism)
- ☐ Political Islam (support)
- ☐ Political Islam (criticism)

Please list any additional topics (one or a few words separated by commas)

Notes or observations: Please type your general observations (1-3 sentences) about the blog or blogger: What stands out most about this blogger?



Arabic Filter

What type of site is this? (add new type by using the field below.)

- ☐ blog
- ☐ blog service
- ☐ news (Web site for offline news, newspaper, radio, etc.)
- ☐ news (online only)
- ☐ portal/directory
- ☐ business
- ☐ individual: politician
- ☐ individual: journalist
- ☐ individual: poet, writer, intellectual
- ☐ individual: other
- ☐ organization: advocacy
- ☐ organization: ngo
- ☐ organization: other
- ☐ political party
- ☐ web 2.0 (social network, user generated content, etc)
- ☐ government
- ☐ online petition
- ☐ Web site educational services
- ☐ picture uploading services
- ☐ Web site that facilitates creating Internet groups
- ☐ collection of variety of links to individual things
- ☐ search engine
- ☐ other

Notes or observations: Please type a short description (1-3 sentences) of the blog, blogger or Web site: What stands out most?

If the site is a blog, please continue. Otherwise click 'complete item' and go to the next site.

What is the nationality of the blogger?

Is the blogger located in his/her country, or an expatriate?

- ☐ in own country
- ☐ expatriate (outside the Arab world)
- ☐ expatriate (in another Arabic-speaking country)
- ☐ don't know

Where are they located specifically (country, city, town, region)? Please write 'dk' (for 'don't know') if you can't tell anything about where they are.

Does the blogger reveal their name, or write anonymously?

- ☐ uses name
- ☐ anonymous (including pseudonyms)

Is the blogger male or female?

- ☐ Don't Know
- ☐ Female
- ☐ Male

Can you estimate the blogger's age? (select the most appropriate category)

- ☐ under 18
- ☐ 18-24 (student age)
- ☐ 25-35 (young adult)
- ☐ 36-60 (adult)
- ☐ over 60 (older adult)
- ☐ don't know

Arabic Outlinks

Is this site (mainly) in Arabic?

- ☐ yes
- ☐ no

What type of site is this?

- ☐ blog (of any type: individual, group, company, etc.)
- ☐ news/info/media (including mainstream media, web-only)
- ☐ social network (like MySpace, Facebook)
- ☐ social media (users contribute media or nominations, like Flickr, YouTube, Digg)
- ☐ portal/search/directory
- ☐ organization (not-for-profit, NGO, etc)
- ☐ government Web site
- ☐ political party/campaign Web site
- ☐ commercial: company site (information about the company)
- ☐ commercial: retail, ecommerce (consumers can buy something)
- ☐ individual person's site (but not a blog, e.g. many book authors, artists have these)
- ☐ educational site (including schools and universities)
- ☐ polling/newswires/data/commercial analysis
- ☐ online discussion forums/bulletin board
- ☐ other

Does the site focus on any of the following topics or issues?

- ☐ politics
- ☐ religion
- ☐ news
- ☐ arts/literature
- ☐ women's issues/rights
- ☐ minority issues/rights
- ☐ pop culture (music, tv, movies)

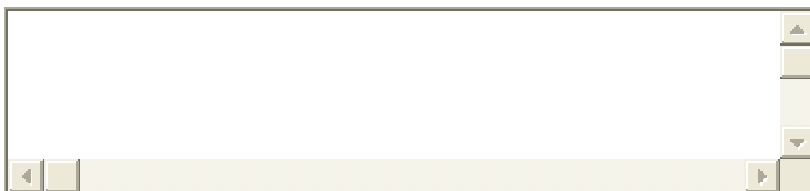
- ☐ sports
- ☐ technology

What country is this site primarily concerned with?

- ☐ Algeria
- ☐ Bahrain
- ☐ Chad
- ☐ Egypt
- ☐ Eritrea
- ☐ Iraq
- ☐ Israel
- ☐ Jordan
- ☐ Kuwait
- ☐ Lebanon
- ☐ Libya
- ☐ Mauritania
- ☐ Morocco
- ☐ Oman
- ☐ Qatar
- ☐ Saudi Arabia
- ☐ Somalia
- ☐ Sudan
- ☐ Syria
- ☐ Tunisia
- ☐ United Arab Emirates
- ☐ United States
- ☐ United Kingdom
- ☐ Yemen
- ☐ Pan-Arabic/Multinational
- ☐ Other

If other, please indicate the country.

Notes or observations: Please type a short description (1-3 sentences) of this site.



6. Wikipedia & networked content analysis

M. Burke and R. Kraut (2008). "Taking up the mop: Identifying future Wikipedia administrators," *Proceedings of the 2008 CHI Conference*, April 5-10, Florence, Italy, 3441-3446.

Taking Up the Mop: Identifying Future Wikipedia Administrators

Moira Burke

Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15217 USA
moira@cmu.edu

Robert Kraut

Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15217 USA
robert.kraut@cmu.edu

Abstract

As Wikipedia grows, so do the messy byproducts of collaboration. Backlogs of administrative work are increasing, suggesting the need for more users with privileged admin status. This paper presents a model of editors who have successfully passed the peer review process to become admins. The lightweight model is based on behavioral metadata and comments, and does not require any page text. It demonstrates that the Wikipedia community has shifted in the last two years to prioritizing policymaking and organization experience over simple article-level coordination, and mere edit count does not lead to adminship. The model can be applied as an AdminFinderBot to automatically search all editors' histories and pick out likely future admins, as a self-evaluation tool, or as a dashboard of relevant statistics for voters evaluating admin candidates.

Keywords

Wikipedia, administrators, management, collaboration

ACM Classification Keywords

H.5.3 [Information Interfaces]: Group and Organization Interfaces - Collaborative computing, Web-based interaction, Computer-supported cooperative work

Introduction

Wikipedia, the collaboratively edited online encyclopedia, reached 2.1 million English articles in December 2007. In the midst of exponential growth of both content and users [7], administrators keep things tidy: they delete copyright violations, protect frequently vandalized pages, block malicious users, move pages when there are name conflicts, exclude bulk vandalism from the recent changes list, and edit the front page. Approximately 1400 editors have successfully passed the rigorous peer review of the Request for Adminship (RfA) process and been given

Copyright is held by the author/owner(s).
CHI 2008, April 5 – April 10, 2008, Florence, Italy
ACM 978-1-60558-012-8/08/04.

administrator privileges, and they are considered trusted custodians of the successful encyclopedia and its community of contributors.

Kittur and colleagues found that while admins once accounted for nearly 60% of editing activity, their influence declined to approximately 10% due to an influx of new editors [7]. Yet admins are working harder than ever: while their edits per month have steadily increased, backlogs of work requiring admin privileges continue to grow¹, suggesting a need for additional editors to become admins.

This paper presents a model of the admin promotion process. The model can be used to identify editors likely to be promoted, as a self-evaluation tool for potential admins, and to provide a dashboard of relevant behavior for RfA voters. The model is lightweight, based on edit counts and comment text available in the public Wikipedia database or the user's contribution page, and not requiring any full text from articles or talk pages. Thus, it can be run quickly to search the entire population of editors, or allow editors to calculate their own likelihood of success on the fly without taxing the server. Social science theory, such as Karau and William's collective effort model [6] suggests that people are more likely to contribute to a collective good such as Wikipedia when they know that they are uniquely qualified for the task, or that the likelihood of success is good. Thus, making strengths and weaknesses in their edit histories salient may prompt editors desiring adminship to behave in ways valued by the community of its admins, or may prompt editors who had not considered applying for adminship to step up. Wikipedia currently has a simple version of this idea; there is a hand-maintained page of users with high edit counts who are not administrators². However, sheer number of edits does not make an RfA nominee likely to pass; "editcountitis" among the voters is frowned upon because some users have high counts from minor changes or not previewing work before publishing.

In addition to the practical benefits to the Wikipedia community, this paper also speaks to long-standing concerns of organizational scholars who have asked what

causes employees to get ahead in their jobs—for example, their experiences and skills, social networks, or job-irrelevant attributes like gender or looks (see [9] for review). Despite protestations that admins are lowly janitors mopping up, in many ways election to admin is a promotion, distinguishing an elite core group from the larger mass of editors. The research described uses policy capture [11] to compare nominally important attributes to those that actually lead to promotion. The behavioral data here is rarely available in conventional organizational settings. Although this paper examines only the influence of job-related experiences on promotion, understanding these influences is a base for any additional analyses.

The Request for Adminship Process

To become an administrator, an editor must undergo a week of scrutiny during which the community builds consensus about the candidate's experience and trustworthiness. Admin tools are not granted lightly; an inexperienced, biased, or ill-intentioned editor could cause significant damage, reducing the encyclopedia's credibility or demotivating other editors. Approximately 2700 editors have been nominated for adminship since 2001 with an overall success rate of 53%. However, the process has gradually grown more rigorous, dropping from a 75.5% success rate through 2005 to 42% in 2006 and 2007, and some early admins have expressed doubt that they would pass muster if their RfA were held today [4]. The process once called "no big deal" by the founder of Wikipedia has become a fairly big deal³. In the Guide to RfAs⁴, the community describes criteria many RfA contributors look for in nominees, including strong edit history, varied experience, and polite interaction with other users. This paper examines which of these criteria are most predictive of success, and which are only nominally used by the community in choosing its admins.

Modeling Successful RfA Candidates

The data include all 1551 Requests for Adminship from January 2006 through October 2007, with 49 RfAs removed

³ <http://en.wikipedia.org/wiki/WP:DEAL>

⁴ <http://en.wikipedia.org/wiki/Wikipedia:GRFA>

¹ http://en.wikipedia.org/wiki/Category:Administrative_backlog

² <http://en.wikipedia.org/wiki/Wikipedia:NA>

for being multiple attempts by the same candidate in one month (all of which failed), bots, sockpuppets (multiple identities held by the same person), or because the nominee's edit history prior to the RfA was not available. An additional 1242 RfAs dating back to 2001 are described in the Conclusion section for changes in criteria over time, but the present model focuses on recent criteria in order to identify future admins. For each RfA, data from the user's contribution history page—up to the month before the RfA—was parsed and counted, and grouped according to the criteria described in the Guide to RfAs. Table 1 provides summary statistics for the features. Features applicable to multiple categories were placed in a single category, and two categories from the Guide to RfAs—trustworthiness and high quality of articles—were excluded from this analysis because they could not be captured from simple edit counts. These two categories are discussed further in the Conclusion.

Strong edit history includes the number of edits to articles and the number of months between the user's first edit and the RfA. Total edit count (Mean=5010.5, Std. Dev=5818.8) is included in the baseline analysis but is replaced by counts in the individual namespaces (e.g. articles, user talk pages) in the final model due to multicollinearity.

Varied experience divides edit counts into individual namespaces, including Wikipedia namespace edits (pages in a subsection focusing mainly on policy, with WikiProjects counted separately), and User pages. The *diversity* metric is a proxy for diverse experience consisting of a count of the number of different areas in which the user has participated, from the set {article, article talk, Wikipedia, Wikipedia talk, user, user talk, articles/categories/templates for deletion (XfD), (un)deletion review, other RfAs, village pump, admin intervention against vandalism (AIV), requests for protection (RfP), admin noticeboard, arbitration committee, mediation committee, and wikiProjects}. So, a user who has edited articles, her own user page, and posted once at the Village Pump would have a diversity score of 3. Actual number of edits in each of these sections is accounted for in the following categories to determine their relative importance.

User interaction includes edits to all talk pages, participation on arbitration or mediation committee pages or wikiquote alerts (an early stage in dispute resolution), posting "welcome" on user talk pages, and including common variants of "please" (including "pls" and "plz") or "thanks" (including "thx") in comment text.

Helping with chores includes reversion of vandalism (noted by "revert" or "rv" in the comment) requesting admin intervention for specific vandals (AIV), requesting protection for a frequently vandalized page, neutrality fixes (noted by "(n)pov" for "neutral point of view"), requesting admin attention (e.g. for inappropriate usernames), and participating in deletion discussions, including articles/categories/templates for deletion (XfD) and (un)deletion reviews. It also includes the percent of the user's total edits marked as minor, which is the designation for spelling or small formatting changes.

Observing consensus includes participation in other editors' RfAs, posting to the Village Pump (a forum for technical and policy discussions), or voting for articles to be deleted or rewritten. Though there are better ways to measure consensus building, they involve natural language processing and thus are prohibitively time-consuming at Wikipedia scale. These measures are lightweight proxies for consensus.

Edit summaries or *comments* include the percent of edits with a human-written comment (edits that only have automatically generated comments are not included), and the average length of comments. Comments are both summaries of changes and conversations between authors, often preempting objections or asking questions of each other [12].

Results

To examine the impact of these behavioral factors on the likelihood of an RfA's success, we first determined baseline performance to be 58.0% for a model that always predicted failure. Adding the nominee's total edit count improved performance to 66.2% accuracy. Then we compared the performance of several models including machine learning approaches using decision trees and support vector machines, and statistical models of regression, applied to all of the features mentioned above. All models performed

approximately the same (between 73% and 75%

accurate), and identified similar behaviors as important so the results of the regression are reported here for ease of presentation.

We performed a probit regression on the binary dependent variable, RfA success. The coefficients in Table 1 represent the change in probability of success when all continuous variables are at their mean values. Results prior to 2006 are reported for comparison. Multiple attempts by the same candidate in a single month were excluded, leaving only one attempt per month, and the candidate's number of previous RfA attempts (in other months) is included as a control variable, showing that each subsequent attempt has a 14.8% lower chance of success than the previous one. All edit count variables with large means have been divided by 1000 in the regression and are noted in the table, but true means are reported. The model is 74.8% accurate in classifying RfA attempts as successful or not.

Strong edit history and varied experience. The results strongly support the notion that mere edit count ("editcountitis") is a poor predictor of Wikipedia adminship. Every 1000 article edits only increases the probability of success by 1.8%. Instead, variety of experience, particularly in the Wikipedia (policy) namespace, is a much better predictor. One Wikipedia policy edit or WikiProject edit is worth ten article edits; one thousand of those edits increases the probability of success by 19.6% and 17.1%, respectively. The diversity score also shows that making a single edit in any additional region of Wikipedia is correlated with a 2.8% increased likelihood of success. Length of experience helps slightly; each additional month between the user's first edit and her RfA helps by 0.4%. Not surprisingly, edits to user pages do not add information to the model, as the norm is to not edit other users' pages except rarely to add rewards known as barnstars. The majority of user page edits would be to these nominees' own user pages, which do not add much value to Wikipedia as a whole.

User interaction. Some forms of user interaction also help; every thousand article talk edits increases the likelihood of success by 6.3%. Article talk pages are mechanisms for coordination and dispute resolution [13], so it is not

surprising that future admins participate more heavily there than do unsuccessful nominees. Somewhat surprisingly, user talk page edits do not affect likelihood of adminship. This is perhaps because the norm is to hold disagreements over content on article talk pages, moving to user talk pages when the disagreement covers multiple articles or a user's overall behavior. Thus, user talk edits are likely to be mixed, and may have higher interpersonal conflict. This is supported by the finding that each edit to an Arbitration or Mediation Committee page, or to a Wikiquote notice, all of which are venues for dispute resolution, decreases the likelihood of success by 0.1%. Politeness helps modestly; though it was rare, every instance of saying "thanks" in a comment increases the likelihood of success by 0.3%. Saying "please" and welcoming newcomers had no effect.

Helping with chores does not have a large impact.

Reverting vandalism is only marginally significant, and so it likely depends on the context of vandalism—such as whether the suspected vandal was an anonymous editor or a well-respected editor with a differing point of view. Pointing out neutrality (pov) issues helped adminship by 0.1%, but escalating problems to the administrators' noticeboard hurt by 0.1%. These results suggest that future administrators stay out of trouble, keeping disputes at the article talk level, rather than bringing them to committees or admin attention. Finally, making minor edits helps, although this may be because inexperienced editors making small changes inflate their importance or do not notice the checkbox for designating the changes as minor, and so it's simply that more responsible editors are promoted.

Observing consensus. Somewhat surprisingly, participating in consensus-building activities, such as other RfAs or the Village Pump, does not increase the likelihood of becoming an admin. However, these are simplistic measures of consensus-building and do not take full text from talk pages into account, which might reveal more successful language patterns of consensus.

Edit summaries/comments. Finally, leaving comments helps. While the length of the comment does not matter, diligently summarizing edits or having brief negotiations with other editors in the comments helps, consistent with

the statement in the Guide to RFAs that some voters expect edit summaries to approach 100%.

Table 1. Descriptive statistics and probit regressions on the likelihood of success in a Request for Adminship in 2006-7 and pre-2006.

	2006-2007				Pre-2006			
	Mean	StDev	Δ	Prob.	Mean	StDev	Δ	Prob.
Attempt number	1.2	0.6	-14.8%	***	1.1	0.3	-11.1%*	
Strong edit history								
Article edits ‡	2611.1	3804.3	1.8%	***	2082.6	2.8	1.1%	
Months since first edit	9.6	8.0	0.4%	*	9.4	7.1	0.2%	
Varied experience								
Wikipedia (policy) edits ‡	433.8	913.8	19.6%	***	265.9	0.5	0.4%	
WikiProject edits ‡	144.0	569.1	17.1%	**	53.0	0.2	7.2%	
Diversity score	11.0	3.7	2.8%	***	9.9	2.6	3.7%***	
User page edits ‡	219.0	296.7	-8.6%		101.8	0.2	11.5%	
User interaction								
Article talk edits ‡	415.2	775.4	6.3%	*	211.6	0.3	15.4%*	
User talk edits ‡	786.6	1169.9	-1.8%		187.3	0.3	-4.8%	
Wikipedia talk edits	87.6	179.5	0%		34.7	64.2	0%	
Arb/mediation/wikiquette edits	9.8	47.1	-0.1%	***	3.7	19.5	-0.2%**	
Newcomer welcomes	76.9	321.1	0%		20.2	90.9	0%	
"Please" in comments	31.7	83.8	0%		20.9	112.5	0%	
"Thanks" in comments	21.8	39.3	0.3%	***	7.2	13.2	0%	
Helping with chores								
"Revert" in comments ‡	257.6	563.2	7.0%	†	98.5	0.2	-6.8%	
Vandal-fighting (AIV) edits	26.5	108.7	0%		0.7	6.5	0.1%	
Requests for protection	3.7	12.2	-0.2%		0.7	2.8	-2.4%***	
"POV" in comments	26.7	46.9	0.1%	**	18.0	28.8	0.0%	
Admin attention/noticeboard edits	18.7	57.9	-0.1%	*	2.9	14.4	0.2%	
X for deletion/review edits ‡	504.5	1027.3	-2.9%		280.7	0.7	2.5%	
Minor edits (%)	25.5	23.0	0.2%	***	35.3	24.3	0.2%***	
Observing consensus								
Other RFAs	27.4	58.9	0%		11.1	30.4	0.0%	
Village pump	9.5	34.6	0%		12.8	56.1	0.0%	
Votes	3.4	15.1	0.1%		1.5	4.0	-0.5%	
Edit summaries / comments								
Commented (%)	63.2	24.0	0.5%	***	75.1	24.8	0.4%***	
Avg. comment length (log2 chars)	5.0	0.8	1.6%		4.8	0.7	3.1%	

N=1502 ***p<.001 **p<.01 *p<.05 †p<0.1 ‡Edit count divided by 1000 for ΔProb

Applications

This model can be applied in three ways: to automatically search all editors' histories and pick out likely future

admins, as a self-evaluation tool, or as a dashboard of relevant statistics for RfA voters.

Following the lessons learned by Cosley and colleagues' SuggestBot [1], which matched pages needing work with editors who had similar interests, a kind of "AdminFinderBot"—a user account for a computer program that runs the model automatically—would need to follow Wikipedia norms and work with the bot approval committee to be most effective and accepted. As a very lightweight process, it already meets one of the main criteria for Wikipedia bots: it would not be a server hog. As the discussion on the non-admin high edit count page shows, some highly contributing editors do not want to become admins, so their privacy will need to be considered in the implementation.

As a self-evaluation tool or voter dashboard, this model would be useful, allowing editors or voters to size up an editor compared to previous RfA nominees, indicating areas where the editor needs improvement, or highlighting the editor's varied experience. Decision-making research shows that people are poor at making these kinds of decisions, and that statistical models often perform better [3][5], so at least providing statistics should improve the RfA process.

Conclusion

Merely performing a lot of production work is insufficient for "promotion" in Wikipedia. Candidates' article edits were weak predictors of success. They also have to demonstrate more managerial behavior. Diverse experience and contributions to the development of policies and WikiProjects were stronger predictors of RfA success. This is consistent with findings that Wikipedia is a bureaucracy [1] and that coordination work has increased substantially [8][13]. However, future work is needed to examine more closely what the admins are doing. Future admins also use article talk pages and comments for coordination and negotiation more often than unsuccessful nominees, and tend to escalate disputes less often.

Although this research has shown that judges pay attention to candidates' job-relevant behavior and especially behavior that suggests the candidate will be a good manager and not just a good worker, it is silent about whether other factors

identified in the organizational literature [9]—social networks, irrelevant attributes, or strategic self-presentation. Indeed, recent evidence that Wikipedia admins use a secret mailing list to coordinate their actions toward others⁵ suggest that sponsorship may also play a role in promotion. Future research in Wikipedia using techniques like those in the current paper can be used to test theories in organizational behavior about criteria for promotion.

An important limitation of the current model is that it does not take the quality of contribution into account. We plan to improve the model by examining measures of length, persistence, and pageviews of edits, which are already being used in more processor intensive models of existing admin behavior [7] and impact of edits [10].

Criteria for admins have changed modestly over time. Success rates were much higher (75.5%) prior to 2006, and collaboration via article talk pages helped more in the past (+15% for every 1000 article talk edits, compared to +6.3% today). The diversity score performs similarly prior to 2006 (+3.7% then, +2.8% now). However, participation in Wikipedia policy and WikiProjects was not predictive of adminship prior to 2006, suggesting the community as a whole is beginning to prioritize policymaking and organization experience over simple article-level coordination.

Acknowledgments

The authors thank Niki Kittur, Tom Murphy, and Ben Collier for feedback. This work was supported by NSF IIS-0325049 and an NSF Graduate Research Fellowship.

References

- [1] Butler, B., Joyce, E., and Pike, J. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. *Proc. CHI 2008*, ACM Press (2008).
- [2] Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. SuggestBot: Using intelligent task routing to help people

find work in Wikipedia. *Proc. IUI 2007*, ACM Press (2007), 32-41.

- [3] Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582.
- [4] Forte, A., and Bruckman, A. Scaling consensus: Increasing decentralization in Wikipedia governance. *Proc. HICSS 2008*.
- [5] Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage Publications.
- [6] Karau, S., and Williams, K. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65 (1993), 681-706.
- [7] Kittur, A., Chi, E., Pendleton, B., Suh, B., and Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Proc CHI 2007*, ACM Press (2007).
- [8] Kittur, A., Suh, B., Pendleton, B. A., Chi, E. (2007). He says, she says: Conflict and coordination in Wikipedia. *Proc CHI 2007*, ACM Press (2007), 453-462.
- [9] Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology*, 58(2), 367-409.
- [10] Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., and Riedl, J. 2007. Creating, destroying, and restoring value in Wikipedia. *Proc GROUP 2007*, ACM Press (2007), 259-268.
- [11] Stumpf, S. A., & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. *The Academy of Management Journal*, 24(4), 752-766.
- [12] Viegas, F., Wattenberg, M., and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. *Proc CHI 2004*, ACM Press (2004), 575-582.
- [13] Viegas, F., Wattenberg, M., Kriss, J., and van Ham, F. Talk before your type: Coordination in Wikipedia. *Proc HICSS 2007*, 575-582.

⁵ http://www.theregister.co.uk/2007/12/04/wikipedia_secret_mailing

T. Chesney (2006). "An empirical examination of Wikipedia's credibility," *First Monday*, 11(11), November.



PEER-REVIEWED JOURNAL ON THE INTERNET

An empirical examination of Wikipedia's credibility

by Thomas Chesney



Wikipedia is an free, online encyclopaedia; anyone can add content or edit existing content. The idea behind Wikipedia is that members of the general public can add their own personal knowledge, anonymously if they wish. Wikipedia then evolves over time into a comprehensive knowledge base on all things. Its popularity has never been questioned, although some have speculated about its authority. By its own admission, Wikipedia contains errors. A number of people have tested Wikipedia's accuracy using destructive methods, *i.e.* deliberately inserting errors. This has been criticised by Wikipedia. This short study examines Wikipedia's credibility by asking 258 research staff with a response rate of 21 percent, to read an article and assess its credibility, the credibility of its author and the credibility of Wikipedia as a whole. Staff were either given an article in their own expert domain or a random article. No difference was found between the two group in terms of their perceived credibility of Wikipedia or of the articles' authors, but a difference was found in the credibility of the articles — the experts found Wikipedia's articles to be more credible than the non-experts. This suggests that the accuracy of Wikipedia is high. However, the results should not be seen as support for Wikipedia as a totally reliable resource as, according to the experts, 13 percent of the articles contain mistakes.

Contents

[Introduction](#)

[Method](#)

[Results](#)

[Discussion](#)

Introduction

Formally launched in 2001 by Jimmy Wales, Wikipedia is an free, online encyclopaedia; anyone can add content or edit existing content. At the time of writing, it contains around 650,000 articles in English and receives about 60 million hits per day [1]. Its popularity has never been questioned, although some have speculated about its authority. The idea behind Wikipedia is that members of the general public can add their own personal knowledge, anonymously if they wish; (a "wiki" refers to a Web page which can be changed by its visitors). Wikipedia then evolves over time into a comprehensive knowledge base on all things. Disputes between authors over content are frequent. These are settled essentially by a process of voting where the authors involved and any other interested parties can discuss the problem between themselves until the truth emerges and is published. Wikipedia claims that none of its articles are ever complete or finished. They also admit that vandalism is a constant problem with people publishing "with an agenda". This can be seen on the pages of controversial subjects such as abortion. Many articles have a message at the top pointing out that the neutrality of the article has been questioned or that the article is being considered for deletion. Each article has a discussion forum attached to it where authors can debate changes, and a history page where everyone can see what changes have been made, by whom and can compare earlier versions.

That anyone can edit articles has led many writers to discredit Wikipedia. Since at any point in time, by Wikipedia's own admission, an article may contain errors, this calls into question the

Proponents of Wikipedia have hit back by pointing out that the site does not claim to be authoritative or reliable but realises that it is "possible for biased, out of date or incorrect information to be posted." [2] They claim that because so many people are contributing, the content will become more reliable as time passes.

[illegible]

So there are two groups, one that is able to assess the content itself and the other which is only able to assess the content on the basis of how it reads. If it is assumed that Wikipedia looks and feels like a professional encyclopaedia, then the second group may be expected to find it reasonably credible. Therefore if Wikipedia, when assessed by the experts, is found not to be credible, it would be expected that the two groups will differ in their responses. Specifically the article's credibility, the writer's credibility and Wikipedia's credibility would be less under Condition 1 than Condition 2. However proponents of Wikipedia may argue that its credibility would be greater under Condition 1 than Condition 2. For this reason, no direction will be specified in the first three hypotheses which are:

H1a Perceived credibility of the author will be different under Condition 1 than Condition 2.

H1b Perceived credibility of the article will be different under Condition 1 than Condition 2.

H1c Perceived credibility of the site will be different under Condition 1 than Condition 2.

It might also be expected that there be a negative correlation between perceived credibility of all respondents and how "cynical" the respondent is of information taken from the Internet. This leads to the following hypotheses:

H2a Perceived credibility of the author will be negatively correlated with how cynical respondents are of Internet information.

H2b Perceived credibility of the article will be negatively correlated with how cynical respondents are of Internet information.

H2c Perceived credibility of the site will be negatively correlated with how cynical respondents are of Internet information.

Measurements

Information with high quality is usually considered to have some or all of the following characteristics: up-to-date, relevant, accurate, economic for the purpose at hand, on time and understandable to the person who needs it. This study measures the credibility of the site, the authors who contribute to the site and the articles they write. Credibility encompasses some but not all of these characteristics. Hovland, *et al.* (1953) operationalised credibility as perceived expertise and perceived trustworthiness. Berlo, *et al.* (1970) suggest three scales to measure credibility: safety (safe-unsafe etc.), qualification (trained-untrained etc.) and dynamism (aggressive-meek etc.). The most consistent dimension of media credibility in studies has been believability (Flanagin and Metzger, 2000). Researchers have also used bias, fairness, accuracy, believability and objectivity (for example Sundar, 1999).

All measurements in this study were made using 7-point Likert scales and were reverse coded if necessary to ensure that low scores represented greater perceptions of credibility and greater cynicism. The author's credibility was measured using nine items taken from Flanagin (2005). The authors were assessed by the extent to which they were perceived to be credible, have high reputation, be successful, be trustworthy, offer information of superior quality, be prestigious, have a sincere interest in important affairs and the extent to which the respondent would be willing to work for them. The mean and standard deviation for this construct were 3.38 and 0.83, and Cronbach's alpha was 0.890.

The article's credibility was measured with five items: believability, accuracy, trustworthiness, bias and completeness. These were taken from Flanagin and Metzger (2000) who adapted them from Austin and Dong (1994), Gaziano (1988), Rimmer and Weaver (1987) and West (1994). The mean and standard deviation for this construct were 4.11 and 1.14, and Cronbach's alpha was 0.838.

Twenty-two items taken from Flanagin (2005) which were adapted from standard source credibility scales (Berlo, *et al.*, 1970; Leathers, 1992; McCroskey, 1966; McCroskey and Jenson, 1975) were used to assess the credibility of the Wikipedia site as a whole. Respondents assessed the extent to which they found the site to be trustworthy, believable, reliable, authoritative, honest, safe, accurate, valuable, informative, professional, attractive, pleasant, colourful, likable, aggressive, involving, bold, interactive, interesting, sophisticated, biased and organized. The mean and standard deviation for this construct were 2.93 and 0.73, and Cronbach's alpha was 0.937.

How cynical each respondent was towards information on the Internet was assessed with five items taken from Metzger, *et al.* (2003). Respondents stated whether they: check to see who the author of a Web site is, check to see if the information is current, seek out other sources to validate information found online, consider whether the information is opinion or fact and check to see that the information is complete and comprehensive. The mean and standard deviation for this construct were 2.77 and 0.90, and Cronbach's alpha was 0.723.

Results

Table 2 shows the mean, standard deviation and standard mean error of the responses for the author, article, site credibility and cynicism of respondents under each condition. Table 3 shows the results of testing for a difference in the means using the t test. Pearson's correlation coefficient was used to examine the relationship between cynicism and credibility. The results of this are shown in Table 4.

Table 2: Mean, standard deviation and standard error of the mean of responses					
		Number	Mean	Standard deviation	Standard error of the mean
<i>Author</i>	Condition 1	30	3.35	0.83	0.15
	Condition 2	24	3.40	0.86	0.18
<i>Article</i>	Condition 1	30	2.66	0.89	0.16
	Condition 2	24	3.14	1.07	0.22
<i>Site</i>	Condition 1	30	2.84	0.62	0.11
	Condition 2	24	3.00	0.85	0.17

Table 3: Results of independent samples T test ($p < 0.1$)			
	t	Sig	
<i>Author credibility</i>	-0.230	0.819	NS
<i>Article credibility</i>	1.797	0.078	S
<i>Site credibility</i>	-0.773	0.433	NS
<i>Cynicism</i>	0.590	0.558	NS

Table 4: Correlation coefficients ($p < 0.05$)		
Correlation between cynicism and ...	r	
<i>Author</i>	-0.036	NS
<i>Article</i>	0.042	NS
<i>Site</i>	-0.105	NS

Discussion

The results show no significant difference between the mean responses under the two conditions when assessing the writer's credibility and the site's (Wikipedia's) credibility. There is a difference between the mean responses when assessing the article's credibility, significant at the 10 percent level. The oddity is the direction of this difference — the experts rated the articles as being more credible than the non-experts. Therefore no support was found for H1a or H1c but support was found for H1b. This difference in the articles' credibility needs to be examined. It may be the case that non-experts are more cynical about information outside of their field and the difference comes from a natural reaction to rate unfamiliar articles as being less credible. If this is correct, it is the credibility of the Condition 2 articles that is being assessed low and not the credibility of the Condition 1 articles that is being assessed high. However, as the cynicism data in Table 3 shows, there was no difference between the groups in terms of their cynicism. However, it should be noted that this cynicism was self-reported and there may be a bias with people wanting to appear more cynical than in fact they are. Having dismissed this as an explanation, that leaves the conclusion that the experts found Wikipedia articles more credible than the non-experts. This suggests that the accuracy of Wikipedia's information is high, although the mean responses in Table 2 suggest perceived

credibility scores of around 3 on a scale from 1 (very credible) to 7 (very incredible). While perhaps not 'high' credibility, this certainly is not 'low'. In the survey, all respondents under Condition 1 were asked if there were any mistakes in the article they had been asked to read. Only five reported seeing mistakes and one of those five reported spelling mistakes rather than factual errors. This suggests that 13 percent of Wikipedia's articles have errors.

No correlation between respondent's cynicism and perceived credibility was found suggesting that perceived credibility is unaffected by how cynical respondents are.

These results may be some cause for cheer by advocates of Wikipedia, but they should be taken with caution. The sample size was small and the difference between the means of the articles' credibility was only significant at the 10 percent level, not at 5 percent. Further work should be done to verify this finding. In any case, the results should not be seen as support for Wikipedia as a totally reliable resource as, according to data collected during this project, 13 percent of the articles contain mistakes. 

About the author

Thomas Chesney is a Lecturer in Information Systems at the Nottingham University Business School.

E-mail: Thomas [dot] Chesney [at] nottingham [dot] ac [dot] uk

Notes

1. <http://en.wikipedia.org/wiki/Wikipedia#History>, accessed 22 August 2005.

2. http://en.wikipedia.org/wiki/Wikipedia:Overview_FAQ, accessed 22 August 2005.

References

E.W. Austin and Q. Dong, 1994. "Source v. content effects on judgments of news believability," *Journalism Quarterly*, volume 71, number 4, pp. 973–984.

D.K. Berlo, J.B. Lemert and R.J. Mertz, 1970. "Dimensions for evaluating the acceptability of message sources," *Public Opinion Quarterly*, volume 33, number 4, pp. 563–576.

S. Burgard, 2005. "Turning over editorial pages to the bloggers is a terrible idea," *The Masthead*, volume 57, number 3, pp. 12–13, and at http://www.findarticles.com/p/articles/mi_qa3771/is_200510/ai_n15640966, accessed 6 November 2006.

B. Cronin, 2005. "Dean's Notes: BLOG: see also Bathetically Ludicrous Online Gibberish," at http://www.slis.indiana.edu/news/story.php?story_id=958, accessed 9 September 2005.

A.J. Flanagin, 2005. "IM online: Instant messaging use among college students," *Communication Research Reports*, volume 22, pp. 175–187.

A.J. Flanagin and M.J. Metzger, 2000. "Credibility of the Internet/World Wide Web," *Journalism and Mass Communication Quarterly*, volume 77, number 3, pp. 515–540.

C. Gaziano, 1988. "How credible is the credibility crisis?" *Journalism Quarterly*, volume 65, number 2, pp. 267–278.

M. Glaser, 2004. "Collaborative Conundrum: Do Wikis Have a Place in the Newsroom?" at <http://ojr.org/ojr/glaser/1094678265.php>, accessed 22 August 2005.

C.I. Hovland, I.L. Janis and H.H. Kelley, 1953. *Communication and persuasion: Psychological studies in opinion change*. New Haven, Conn.: Yale University Press, pp. 19–48.

D.G. Leathers, 1992. *Successful nonverbal communication: Principles and applications*. Second edition. New York: Macmillan.

P. Leppik 2004. "How Authoritative is Wikipedia" at <http://www.frozenorth.org/C2011481421/E652809545/index.html>, accessed 22 August 2005.

A. Martinez, 2005. "Innovations bring transparency that editorial pages need," *The Masthead*, volume 57, number 3, pp. 4–5, and at http://www.findarticles.com/p/articles/mi_qa3771/is_200510/ai_n15640944, accessed 6 November 2006.

M.J. Metzger, A.J. Flanagin and L. Zwarun, 2003. "College student Web use, perceptions of information credibility, and verification behavior," *Computers & Education*, volume 41, pp. 271–290.

J.C. McCroskey, 1966. "Scales for the measurement of ethos," *Speech Monographs*, volume 33, pp. 65–72.

J.C. McCroskey and T.A. Jenson, 1975. "Image of mass media news sources," *Journal of Broadcasting*, volume 19 pp. 169–180.

T. Rimmer and D. Weaver, 1987. "Different questions, different answers? Media use and media credibility," *Journalism Quarterly*, volume 64, pp. 28–36.

K.G. Schneider, 2005. "Wikipedia," at <http://freerangelibrarian.com/archives/052905/wikipedia.php>, accessed 22 August 2005.

S.S. Sundar, 1999. "Exploring receivers' criteria for perception of print and online news," *Journalism and Mass Communication Quarterly*, volume 76, number 2 pp. 373–387.

M.D. West, 1994. "Validating a scale for the measurement of credibility: A covariance structure modeling approach," *Journalism and Mass Communication Quarterly*, volume 71, number 1, pp. 159–168.

Editorial history

Paper received 16 May 2006; accepted 24 August 2006.

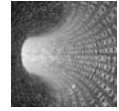
Contents **Index**



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 2.0 License](http://creativecommons.org/licenses/by-nc-sa/2.0/).

An empirical examination of Wikipedia's credibility by Thomas Chesney
First Monday, volume 11, number 11 (November 2006),
URL: http://firstmonday.org/issues/issue11_11/chesney/index.html

S. Niederer and J. van Dijck (2009).
"The case of Wikipedia: Wisdom of the
crowd or technicity of content?"
unpublished ms.



Article

Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system

new media & society

XX(X) 1–20

© The Author(s) 2010

Reprints and permission: sagepub.

co.uk/journalsPermissions.nav

DOI: 10.1177/1461444810365297

<http://nms.sagepub.com>



Sabine Niederer

University of Amsterdam, The Netherlands

José van Dijck

University of Amsterdam, The Netherlands

Abstract

Wikipedia is often considered as an example of ‘collaborative knowledge’. Researchers have contested the value of Wikipedia content on various accounts. Some have disputed the ability of anonymous amateurs to produce quality information, while others have contested Wikipedia’s claim to accuracy and neutrality. Even if these concerns about Wikipedia as an encyclopaedic genre are relevant, they misguidedly focus on human agents only. Wikipedia’s advance is not only enabled by its human resources, but is equally defined by the technological tools and managerial dynamics that structure and maintain its content. This article analyses the sociotechnical system – the intricate collaboration between human users and automated content agents – that defines Wikipedia as a knowledge instrument.

Key words

collaborative knowledge, protocol, sociotechnical system, Web 2.0, Wikipedia

Introduction

User-generated content has revived the idea of the web as a place of human collaboration and a place of activity for ‘everybody’ (Shirky, 2008). The Wikipedia project is often considered as an example par excellence of ‘collaborative knowledge’, of ‘social media’

Corresponding author:

Sabine Niederer, University of Amsterdam, Turfdraagsterpad 9, 1012 XT Amsterdam, The Netherlands.

Email: s.m.c.niederer@uva.nl

or the ‘wisdom of crowds’. Since 2001, a group of editors and volunteers have engaged in developing an online encyclopaedia, whereby everyone is invited to contribute and articles are open to continuous editing. Large numbers of contributors, the so-called ‘Wikipedians’, produce an online encyclopaedia that is unprecedented in scale and scope. Researchers who have evaluated or contested the value of Wikipedia content have almost unanimously focused on its human contributors. For instance, it has been disputed whether an encyclopaedia that is produced by many (anonymous) minds results in quality information (Keen, 2008). Other critics have contested Wikipedia’s claim to accuracy and neutrality by pointing at the liability of allowing anonymous contributors whose interest or expertise remains undisclosed.

Even if these concerns about Wikipedia as an encyclopaedic genre are legitimate and relevant, we argue that they misguidedly focus on human agents only, while neglecting the role of technology. In this article, we focus on the intricate interrelation between human and technological tools which lies at the heart of several debates concerning Wikipedia. The first debate revolves around the question of whether Wikipedia is authored primarily by a few elite users or by many common contributors, an opposition we would like to question, for Wikipedia has a refined hierarchical structure in which contributing administrators, registered users, anonymous users and ‘bots’ (short for ‘software robots’; see below) all have a distinct rank in an orderly system. Second, we would like to refocus the public debate on the quality of Wikipedia’s encyclopaedic information (disputing whether its entries are accurate and neutral) by shifting attention to the protocols and technologies deployed to facilitate consensus editing. A basic comprehension of Wikipedia’s automated editing systems, as well as emerging tracking tools like the WikiScanner, are needed to evaluate the encyclopaedia’s ability to meet standards of neutrality and accuracy, while preventing overt bias and vandalism.

In the third section of this article, we want to show how dependent various user groups and entries are on non-human content agents (or bots) that assist in editing Wikipedia content. Examining the variable dependency of human editors on bots for editing encyclopaedic content per language Wikipedia, we explore how an automated system of ‘interwiki’ and ‘interlanguage’ bots helps maintain the overall content strategies of the online encyclopaedia. Linking and networking specific language Wikipedias into one global system is less the result of people working together across languages and borders as it is the product of collaboration between humans and bots.

From our analysis, we conclude that any evaluation of Wikipedia’s qualities should acknowledge the significance of the encyclopaedia’s dynamic nature as well as the power of its partially automated content management system. It is the intricate collaboration between large numbers of human users and sophisticated automated systems that defines Wikipedia’s ultimate success as a knowledge instrument. In order to unravel this intricate human-technological interaction, we deploy several ‘natively’ digital methods for web research (Rogers, 2009).¹ In his publication *The End of the Virtual*, Rogers calls for ‘research *with* the Internet’ that applies novel medium-specific methods and tools. Rather than importing existing analytical methods onto the web, natively digital methods can ‘move beyond the study of online culture alone’ (2009: 5) and help understand new media as the interplay between human and technological agents. The concrete goal of this analysis is to provide a better understanding of how Wikipedia’s automated

technological systems and the management of large numbers of edits are inextricably intertwined. More philosophically, we want to theorize human and machine contributions as complementary parts of a sociotechnical system that lies at the heart of many Web 2.0 platforms.

Many minds collaborating

Wikipedia has been described with terms such as ‘many minds’ (Sunstein, 2006) and similar notions such as ‘the wisdom of crowds’ (Kittur and Kraut, 2008; Surowiecki, 2004), ‘distributed collaboration’ (Shirky, 2008), ‘mass collaboration’ (Tapscott and Williams, 2006), ‘produsage’ (Bruns, 2008), ‘crowdsourcing’ (Howe, 2006), ‘Open Source Intelligence’ (Stalder and Hirsh, 2002) and ‘collaborative knowledge’ (Poe, 2006). The collectively written encyclopaedia on a wiki platform is often heralded as an example of collaborative knowledge production at its best. In early 2008, an article in the *New York Review of Books* explained the compelling charm of Wikipedia:

So there was this exhilarating sense of mission – of proving the greatness of the Internet through an unheard-of collaboration. Very smart people dropped other pursuits and spent days and weeks and sometimes years of their lives doing ‘stub dumps,’ writing ancillary software, categorizing and linking topics, making and remaking and smoothing out articles – without getting any recognition except for the occasional congratulatory barnstar on their user page and the satisfaction of secret fame. Wikipedia flourished partly because it was a shrine to altruism – a place for shy, learned people to deposit their trawls. (Baker, 2008)

Since the start of the Wikipedia project in 2001, the dedication of its contributors, as well as the group effort as an alternative to the professional expert approach, have been sources of both excitement and criticism. Even if Wikipedia has now become famous for its collaborative character of many minds producing knowledge, it is interesting to remind ourselves that the project originally intended to be an expert-generated encyclopaedia. Started with the name of ‘Nupedia’, a small team of selected academics was invited to write the entries, with the aim of creating a ‘free online encyclopaedia of high quality’ (Shirky, 2008: 109). The articles would be made available with an open content licence. Founder Jimmy ‘Jimbo’ Wales and his employee Larry Sanger put into place a protocol based on academic peer review (Poe, 2006; Shirky, 2008). This expert approach failed partly because of the slowness of the editing process by invited scholars. To speed up the process, Sanger suggested a wiki as a collective place where scholars and interested laypeople from all over the globe could help with publishing and editing draft articles. The success of Wikipedia and the commitment of the Wikipedians took them by surprise. Sanger became the ‘chief organizer’, a wiki-friendly alternative to the job of ‘editor-in-chief’ that he held for Nupedia. He made a great effort to keep Wikipedia organized, while at the same time providing space for some of the ‘messiness’ (edit wars, inaccuracies, mistakes, fights, and so on) that collaborative editing brings along. In early 2002, however, Sanger turned away from the epistemic free-for-all of Wikipedia towards an expert-written encyclopaedic model called Citizendium (http://en.citizendium.org/wiki/Welcome_to_Citizendium), while Wales chose to further pursue the Wiki model.

The question as to whether online encyclopaedias and similar enterprises *should be* produced by few (expert) or many (amateur) minds has been the source of heated debate ever since the Sanger/Wales split. Internet critic Andrew Keen (2008: 186) applauded Sanger for coming to his senses about the debased value of amateur contributions in favour of expert professionals. On the other end of the spectrum, many Wikipedia fans have praised its democratizing potential as well as its ethos of community and collaboration, a source of knowledge free for everyone to read and write (Benkler, 2006; Jenkins, 2006). By the same token, the notion that Wikipedia is actually produced by ‘crowds’ has been regularly challenged, most notably by Wikipedia’s founders. During the first five years of its existence, Wikipedia was largely dependent upon the work of a small group of dedicated volunteers. Although they soon formed a thriving community, the notion of a massive collective of contributors was repeatedly downplayed by Wales. As he pointed out in a talk at Stanford University in 2006:

The idea that a lot of people have of Wikipedia is that it’s some emergent phenomenon – the wisdom of mobs, swarm intelligence, that sort of thing – thousands and thousands of individual users each adding a little bit of content and out of this emerges a coherent body of work (...) [But Wikipedia is in fact written by] a community, a dedicated group of a few hundred volunteers (...) I expected to find something like an 80–20 rule: 80% of the work being done by 20% of the users (...) But it’s actually much, much tighter than that: it turns out over 50% of all the edits are done by just [0].7% of the users. (Wales cited in Swartz, 2006)

As Wales asserts, until 2006, Wikipedia was largely written and maintained by a small core of dedicated editors (2% doing 73.4% of all the edits). The disproportionate contribution of (self-)designated developers versus ‘common users’ can also be found in research into the open source movement. Rishab Aiyer Ghosh and Vipul Ved Prakash were among the first to disaggregate the ‘many minds’ collaborating in the open software movement. Their conclusion was that, ‘free software development is less a bazaar of several developers involved in several projects and more a collation of projects developed single-mindedly by a large number of authors’ (Ghosh and Prakash, 2000: 1). In the open source movement, very few people were actually collaborating in developing software.

It would be a mistake, however, to dismiss the idea of Wikipedia’s ‘many contributors’ as a myth. Starting in 2006, the online encyclopaedia showed a distinct decline in ‘elite’ users, while, at the same time, the number of edits made by novice users and ‘masses’ was steadily increasing. Various researchers have pointed to the dramatic shift in the workload to the common user (Kittur et al., 2008). Instead of pitching the power of the expert versus the wisdom of the crowds, Kittur et al. speak of ‘the rise of the bourgeoisie’, a marked growth in the population of low-edit users between 2006 and 2008. Interestingly, these researchers explain this shift by describing Wikipedia in terms of a dynamic social system that evolves as a result of the gradual development, implementation and distribution of content management systems. After an initial period of being managed by a small group of high-powered, dedicated volunteers, the ‘pioneers were dwarfed by the influx of settlers’ (Kittur et al., 2008: 8). The early adopters select and refine technological and managerial systems, followed by a majority of novice users who

begin to be the primary users of the system. Kittur and his colleagues observe a similar decline in elite users of Web 2.0 platforms and suggest that it may be a common phenomenon in the evolution of online collaborative knowledge systems. This tentative conclusion is underscored by other researchers who show that, in order to sustain the encyclopaedia’s growing popularity, its organizers need to identify more productive workers and grant them ‘administrator’s status’ (Burke and Kraut, 2008).

Although these researchers correctly observe significant changes in the ‘wisdom of crowds’ paradigm, they seem to be stuck in the antagonism of (few) experts versus (many) common users. Even if they notice the growing presence of non-human actors in the evolution of Wikipedia’s social dynamics, such as software tools and managerial protocols, they tend to underestimate their importance. In fact, the increasing openness of Wikipedia to inexperienced users is *made possible* by a sophisticated technomanagerial system, which facilitates collaboration on various levels. Without the implementation of this strict hierarchical content management system, Wikipedia would most likely have become a chaotic experiment.

According to Alexander Galloway, the internet and many of its (open source) applications are not simply open or closed, but *modulated*. Networked technology and management style are moderated by protocol, which gains its authority ‘from technology itself and how people program it’ (2004: 121). Wikipedia, built as an open system and carried out by large numbers of contributors, appears to be a ‘warm, friendly’ technological space, but only becomes warm and friendly through what Galloway refers to as ‘technical standardization, agreement, organized implementation, broad adoption and directed participation’ (2004: 142).

This is exactly what happened during the first five years of Wikipedia, during which time administrators developed strict protocols for distributing permission levels, imposing a hierarchical order in deciding what entries to include or exclude, what edits to allow or block. If we look more closely at Wikipedia’s organizational hierarchy (see Figure 1),

Permission level	Wikipedia users
Most permissions	Developer/System administrator
	Steward
	Check user
	Oversight
	Bureaucrat
	Administrator/Sysop
	Bot
	Registered user
	Newly registered user
	Anonymous user
No permissions	Blocked user

Figure 1. Schematic overview of global and local categories of Wikipedia users according to permission levels. Available at: http://meta.wikimedia.org/wiki/User_groups

we can distinguish various user groups, some of which are ‘global’ (in the sense that they edit across various language Wikipedias) while others are specific to a certain ‘local’ Wikipedia.

Each user group maintains the same pecking order, regulating the distribution of permission levels: ‘blocked users’ have the least permissions, for they can only edit their own talk page; anonymous users have fewer permissions than registered users, who, in turn, are at a lower level of permission than bots; bots are just below administrators (‘admins’), who occupy the highest level in the elaborate Wikipedia bureaucracy; system administrators (or developers) have the most permissions, including server access. This is a small user group of only 10 people who ‘manage and maintain the Wikimedia Foundation Servers’ (http://meta.wikimedia.org/wiki/System_administrators). Remarkable in this ranking system is the position of bots, whose permission level is just below that of administrators, but well above the authority of registered users. We will return to the status of bots in the third section. For now, it is important to note the significant role of automated mechanisms in the control of content.

In fact, we could argue that the very success of the Wikipedia project lies in the regulation of collaborative production at *any level*, from a small edit or a single upload to a more extensive contribution or even development of the platform or its content. Like any large public system, Wikipedia works through a system of disciplinary control by issuing rewards, such as granting a dedicated user the authority level of administrator (Burke and Kraut, 2008) and by blocking contributors’ rights to those users who deviate from the rules. A disciplinary system of power distribution in the digital age, however, cannot be regarded exclusively as a system of *social* control. As Gilles Deleuze (1990) has pointed out in his acute revision of Foucault’s disciplinary institutions, a ‘society of control’ deploys technology as an intricate part of its social mechanisms. Wikipedia’s content management system, with distinct levels of permissions, allows moreover for *protocological control*, a mode of control that is at once social and technological – one cannot exist without the other (Galloway, 2004: 17). Along the same lines, Bruno Latour (1991: 129) proposes to analyse technological objects and infrastructures as ‘sociotechnical ensembles’, in which the strict division between ‘material infrastructure’ and ‘social superstructure’ is to be dissolved:

Rather than asking ‘is this social’ or ‘is this technical or scientific’ ... we simply ask: has a human replaced a non-human? Has a non-human replaced a human? ... Power is not a property of any of those elements [of humans or non-humans] but of a chain. (1991: 110)

Similar to Latour’s attempt to dissolve the ‘technology/society’ divide, we argue that the dynamic interwovenness of human and non-human content agents is an underrated yet crucial aspect of Wikipedia’s performance. The online encyclopaedia’s success is based on sociotechnical protocological control, a combination of its technical infrastructure and the collective ‘wisdom’ of its contributors. Rather than assessing Wikipedia’s epistemology exclusively in terms of ‘power of the few’ versus the ‘wisdom of crowds’, we propose to define Wikipedia as a gradually evolving sociotechnical system that carefully orchestrates all kinds of human and non-human contributors by implementing managerial hierarchies, protocols and automated editing systems.

Accurate and neutral encyclopaedic information

A similar disregard of technological aspects can be observed in another heated debate that has haunted the online project from its inception: the question regarding the quality of Wikipedia's encyclopaedic information. Wikipedia entries have often been held against the standards of *accuracy* and *objectivity* set by reputed encyclopaedias such as the Encyclopaedia Britannica. Wikipedia entries are based on three core principles, which serve as leading rules for its contributors and aim at holding up the encyclopaedia's quality standards.

The first core rule is 'verifiability' (i.e. readers have to be able to retrieve Wikipedia content in reliable sources). Therefore, referring to published articles and verifiable resources is necessary to have the article (or edits) accepted (<http://en.wikipedia.org/wiki/Wikipedia:Verifiability>). A second, related core rule is called 'no original research'. Wikipedia simply does not accept 'new' (unpublished) research or thought (http://en.wikipedia.org/wiki/Wikipedia:No_original_research). Again, reliability on Wikipedia means citing proven published sources. Third, articles have to be written from a 'neutral point of view' (NPOV) to avoid bias, meaning the articles have to be based on facts, and facts about opinions, but not on opinions (http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view). All contributors, whether single anonymous users or administrators, are required to comply with these rules.² Compliance is regulated by the abovementioned core rules, and non-compliance is punished by removal of edits. In past debates on Wikipedia's standards of accuracy and neutrality, the emphasis has been on whether they can be kept up by crowds of human users. The more profound question in line with our research thesis is, however, how these standards are maintained and controlled through the organization and mechanics of Wikipedia's content management system.

Initially, the quality debate concentrated mainly on accuracy or, more precisely, on the lack thereof due to the impossibility of verifying and authenticating sources. With so many anonymous and amateur contributors, the likeliness of vandalism, inaccuracy and downright sloppiness in factual details was more than real. As danah boyd (2005) observes, Wikipedia 'lacks the necessary research and precision' and 'students are often not media-savvy enough to recognize when to trust Wikipedia and when this is a dreadful idea'. Other researchers entered the quality of content debate by testing Wikipedia's robustness in terms of content vandalism. Alexander Halavais (2004) intentionally contributed incorrect information to existing articles. For his 'Isuzu experiment', he inserted 13 mistakes into 13 different articles, expecting that most of the errors would remain intact. Much to Halavais's surprise, his wrongful edits did not last long, but were all corrected within a couple of hours.³

However, the explanation for the speed at which his vandalism was detected lies less in human acuity than in technological perspicacity. The fact that Halavais had made all his changes from the same username and IP address arguably made it all too easy for Wikipedians and their tools to undo his edits. Making 13 changes in 13 different articles in a short timeframe obviously attracts attention from automated bots, and even human Wikipedians, after spotting one mistake, would have probably looked into his other edits in the other articles and could have easily retrieved the other mistakes 'by association'. Philosopher of science P.D. Magnus therefore provided a corrective to Halavais's

research method in his 2008 study, in which he inserted inaccuracies distributed across different IP addresses and fields of expertise. He found that one-third of the errors were corrected within 48 hours and that most of the others were 'corrected by association', as was the case with Halavais's experiment, whereby Wikipedians probably started checking his other edits after initially finding three mistakes. Some researchers conclude from these tests that the online encyclopaedia is robust to vandalism due to its huge numbers of watchful community members (Poe, 2006). Instead, we argue that it is rather the strict implementation of protocological control and the use of automated bots that account for Wikipedia's vigilance. With their experiments, Halavais and Magnus have less proven the reliability of Wikipedia articles than the reliability of the encyclopaedia's techno-managerial system.

The notion of a community of vigilant users has continued to feed the accuracy debate, particularly by academic researchers who questioned the reliability of Wikipedia's supposed egalitarian approach. Collaboratively written amateur content, as these critics contend, finds itself at odds with knowledge production. If not written by known experts, how accurate and reliable are these encyclopaedic entries? In December 2005, the first academic research that systematically compared the accuracy of Wikipedia and Encyclopaedia Britannica was published in *Nature* (Giles, 2005). Researchers compared the two encyclopaedias by checking 42 science articles in both publications. The reviewers were academics, who checked the articles without knowing their source. They found Wikipedia and Britannica to be almost equally accurate; not surprisingly, the news was triumphantly announced as 'Wikipedia Survives Research Test' (BBC, 2005). With this outcome, Wikipedia was recognized as an encyclopaedia, at least on the level of its accuracy. But this was not the symbolic end of the reliability discussion. On the contrary, the debate heated up and more research followed. In 2006, information systems researcher Thomas Chesney (2006) conducted more empirical research into the credibility of Wikipedia, asking a total of 258 experts (academics) and non-experts to fill out a survey about a Wikipedia article from their area of expertise (or, for the laymen, in their realm of interest). The respondents found mistakes in 13 percent of the Wikipedia articles. But Chesney also found that the experts gave the Wikipedia articles a higher credibility rating than did the non-experts. Contrary to what Sanger described as the 'perceived inaccuracy of Wikipedia', the respondents expected (and found) Wikipedia to be a reliable source of information on the web.

In response to this accuracy debate, centring on the assumed polarity between (known) experts and (unknown) laypersons, few academics proposed to redirect its focus from *product* to *process* and from the abilities of *people* to the qualities of its technological *tools*. Historian Roy Rosenzweig (2006), who conducted a thorough analysis of Wikipedia biographical entries and compared them to entries from the American National Biography Online (written by known scholars), concludes that the value of Wikipedia should not be sought in the accuracy of its published content at one moment in time, but in the dynamics of its continuous editing process – an intricate process whereby amateurs and experts collaborate in an extremely disciplined manner to improve entries each time they are edited. Rosenzweig notices the benefits of many edits to the factuality of an entry. As he points out, it is not so much crowds of anonymous users that make Wikipedia a reliable resource, but a regulated system of consensus editing that bares *how* history is

written: 'Although Wikipedia as a product is problematic as a sole source of information, the process of creating Wikipedia fosters an appreciation of the very skills that historians try to teach' (2006: 138). One of the most important features, in this respect, is the website's built-in history page for each article, which lets you check the edit history of an entry. According to Rosenzweig, the history of an article as well as personal watch lists and recent changes pages are important instruments that give users additional clues to determine the quality of individual Wikipedia entries.

Part of the discussion disputing the accuracy and neutrality of Wikipedia's content concentrated on the inherent unreliability of *anonymous* sources. How can an entry be neutral and objective if the encyclopaedia accepts copy edits from anonymous contributors who might have a vested interest in its outcome? Critics like Keen (2008) and Denning et al. (2005) have objected to the distribution of editing rights to all users. What remains unsaid in this debate is that the impact of anonymous contributors is clearly restricted due to technological and protocological control mechanisms. For one thing, every erroneous anonymous edit is systematically overruled by anyone who has a (similar or) higher level of permission (which is anyone except for blocked users). Since anonymous users are very low in the Wikipedia pecking order, their edit longevity is likely to be short when they break the rules of objectivity and neutrality.

On top of that, there is an increasing availability of 'counter tools' that allow for checking the identity of contributors or at least their location of origin. On the history page of each Wikipedia entry, we can find the time stamp and IP address for every anonymous edit made. The WikiScanner, a tool created by California Institute of Technology student Virgil Griffith in 2007, makes it possible to geo-locate anonymous edits by looking up the IP addresses in an IP-to-Geo database, listing the IP addresses and the companies and institutions they belong to. It facilitates the tracking of anonymous users by revealing who and where they actually are. The WikiScanner has proven to be a powerful tool for journalists trying to localize and expose biased content. In the WikiScanner FAQ on his website, Griffith states that he created the WikiScanner (among other reasons) to 'create a fireworks display of public relations disasters in which everyone brings their own fireworks, and enjoys' (2008a). The WikiScanner was designed to reveal bias, and Griffith collects the most spectacular results on his website.⁴

The debates concerning Wikipedia's accuracy and neutrality have been dominated by fallacious oppositions of human actors (experts versus amateurs, registered versus anonymous users) and have favoured a static evaluation of its content (correct or incorrect at one particular moment in time). Both qualifications, however, are ill suited when applied to a dynamic online encyclopaedia such as Wikipedia, mostly because a debate grounded in such parameters fails to acknowledge the crucial impact of a non-human actor: Wikipedia's dynamic content management system and the protocols by which it is run. Arguably, Wikipedia is neither the often-advertised platform for many minds, nor is it a space for anonymous knowledge production. The WikiScanner has made the revealing of anonymous users much easier by matching IP addresses with contact information.⁵ Bias can now be identified, tracked and, if necessary, reverted. But there is more to the technicity of Wikipedia content than fast users armed with notification feeds and monitoring devices. The technicity of Wikipedia content, as we will show in the next section, lies in the totality of tools and software robots used for creating, editing and linking

entries, combating vandalism, banning users, scraping and feeding content and cleaning articles. It is the complex collaboration not of crowds, but of human *and* non-human agents combined that defines the quality standards of Wikipedia content.

Co-authored by bots

The significant presence of bots appears counter to the common assumption that Wikipedia is authored by human ‘crowds’. In fact, human editors would never be able to keep up the online encyclopaedia if they were not assisted by a large number of software robots. Bots are pieces of software or scripts that are designed to ‘make automated edits without the necessity of human decision-making’ (http://en.wikipedia.org/wiki/Wikipedia:Bot_policy). They can be recognized by a username that contains the word ‘bot’, such as SieBot or TxiKiBoT. Bots are created by Wikipedians and, once approved, they obtain their own user page and form their own user group with a certain level of access and administrative rights, made visible by flags on a user account page. One year after Wikipedia was founded, bots were introduced as useful helpers for repetitive administrative tasks (http://en.wikipedia.org/wiki/Wikipedia:History_of_Wikipedia_bots). Since the first bot was created on Wikipedia, the number of bots has grown exponentially. In 2002, there was only one active bot on Wikipedia; in 2006, the number had grown to 151 and, in 2008, there were 457 active bots (http://en.wikipedia.org/wiki/Wikipedia:Editing_frequency/All_bots).

In general, there are two types of bots: editing (or ‘co-authoring’) bots and non-editing (or administrative) bots. Each of the bots has a very specific approach to Wikipedia content, related to its often narrow task. Administrative bots are most well known and well liked among Wikipedia users. They are deployed to perform policing tasks, such as blocking spam and detecting vandalism. Vandalism combat bots come into action when ‘vandalism-like’ edits are made. Vandalism is recognizable, for it often means a large amount of deleted content in an article or a ‘more than usual’ change in content. Spellchecking bots check language and make corrections in Wikipedia articles. Ban enforcement bots can block a user from Wikipedia and, thus, take away his or her editing rights, which is something a registered user is not able to do. Non-editing bots are also data miners, used to extract information from Wikipedia, and copyright violation identifiers; the latter compare text in new Wikipedia entries to what is already available on the web about that specific topic and report this to a page for human editors to review. Most bots, being created to perform repetitive tasks, make many edits. In 2004, the first bots reached the record number of 100,000 edits.

The second category of editing or co-authoring bots seems to be much less known by Wikipedia users and researchers (for it would otherwise certainly have played a role in the debates about reliability and accuracy). While not every bot is an author, all bots can be classified as ‘content agents’, as they all actively engage with Wikipedia content. The most active Wikipedians are in fact bots. A closer look at various user groups reveals that bots create a large number of revisions of high quality (http://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits#List). Adler et al. (2008) discovered that the two highest contributors in their ‘edit longevity survival test’ were bots. As mentioned before, bots as a user group have more rights than registered users and also a very specific set of permissions. For instance, bot edits are by default invisible in recent changes logs and watch lists. Research cited above has

already pointed out that Wikipedians rely on these notification systems and feeds for the upkeep of articles.

Describing Wikipedians in bipolar categories of humans and non-humans, however, does not do justice to what is, in fact, a third category: that of the many active users assisted by administrative and monitoring tools, also referred to as software-assisted human editors. Bots are Wikipedians' co-authors of many entries. One of the first editing bots to be deployed by Wikipedians was rambot, a piece of software created by Derek Ramsey (<http://en.wikipedia.org/wiki/User:Ram-Man>). Rambot pulls content from public databases and feeds it into Wikipedia, creating or editing articles on specific content, either one by one or as a batch. Since its inception in 2002, rambot has created approximately 30,000 articles on US cities and counties on Wikipedia using data from the CIA World Factbook and the US census. And since content produced by authoring bots relies heavily on their sources, errors in the data set caused rambot to publish around 2000 corrupted articles. In the course of time, bot-generated articles on American cities and counties were corrected and complemented by human editors, following a strict format protocol: history, geography, demographics, and so on. The articles appear strikingly tidy and informative and remarkably uniform. If we compare, for instance, an article on La Grange, Illinois, as created by rambot in 2002, with a more recent version of this article from 2009, it clearly shows the outcomes of a collaborative editing process; the entry has been enriched with facts, figures and images (see Figure 2). The basic format, however, has remained the same. To date, it is still rambot's main task to create and edit articles about US counties and cities, while human editors check and complement the facts provided by this software robot.

But how dependent is Wikipedia on the use of bots for the creation and editing of content? What is the relative balance of human versus non-human contributions in the online encyclopaedia? Peculiarly, the answer to this simple question turns out to be layered and nuanced. When we started to look for answers, we found there to be striking differences between various language Wikipedias. As a global project, Wikipedia features over 10 million articles in over 250 languages (2 million in English) serving a large number of language communities.⁶ The fact that Wikipedia distinguishes between local and global user groups already suggests that bot activity might differ across local Wikipedias, which indeed turns out to be the case.⁷ Specific language Wikipedias not only greatly vary in size and number of articles, but also in bot activity. The Wikipedia Bot Activity Matrix, Wikipedia's own meta-data record, offers an overview of total bot activity as well as bot activity per language (<http://stats.wikimedia.org/EN/BotActivityMatrix.htm>). The percentage of bot edits in all Wikipedias combined is 21.5 per cent. Excluding the English language Wikipedia, total bot activity amounts to 39 per cent, which means that bot activity is unevenly deployed across languages and communities.

In order to account for the differences in bot activity versus human activity, it is interesting to compare bot activity in the most used language Wikipedias (English, Japanese, German) to bot activity in endangered and revived language Wikipedias (such as Cornish, Oriya, Ladino). In the Dorling maps (see Figures 3 and 4), the thin-lined outer circle depicts the language Wikipedia, sized according to the amount of articles in total. The inner dot represents the share of bot activity in that language Wikipedia. The English, Japanese and German Wikipedias show that by far most of its editing is done by human editors. The German Wikipedia, for instance, has only 9 per cent bot activity; the English

La Grange, Illinois

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by Ram-Man (talk | contribs) at 17:55, 18 October 2002. It may differ significantly from the current revision.
(diff) ← Previous revision | Current revision (diff) | Newer revision → (diff)

La Grange is a village located in Cook County, Illinois. As of the 2000 census, the village had a total population of 15,608.

Geography

According to the United States Census Bureau, the village has a total area of 6.5 km² (2.5 mi²), 6.5 km² (2.5 mi²) of it is land and 0.0 km² (0.0 mi²) of it is water. The total area is 0.00% water.

Demographics

As of the census of 2000, there are 15,608 people, 5,624 households, and 4,049 families residing in the village. The population density is 2,400.9/km² (6,220.7/mi²). There are 5,781 housing units at an average density of 889.3 persons/km² (2,304.1 persons/mi²). The racial makeup of the village is 91.02% White, 6.02% African American, 0.09% Native American, 1.00% Asian, 0.02% Pacific Islander, 0.99% from other races, and 0.86% from two or more races. 3.68% of the population are Hispanic or Latino of any race. There are 5,624 households out of which 37.9% have children under the age of 18 living with them, 60.3% are married couples living together, 9.2% have a woman whose husband does not live with her, and 28.0% are non-families. 24.5% of all households are made up of individuals and 9.7% have someone living alone who is 65 years of age or older. The average household size is 2.67 and the average family size is 3.23. In the village the population is spread out with 28.5% under the age of 18, 4.8% from 18 to 24, 29.4% from 25 to 44, 23.8% from 45 to 64, and 13.5% who are 65 years of age or older. The median age is 38 years. For every 100 females there are 94.2 males. For every 100 females age 18 and over, there are 87.4 males. The median income for a household in the village is \$90,342, and the median income for a family is \$95,554. Males have a median income of \$62,030 versus \$41,260 for females. The per capita income for the village is \$34,887. 4.0% of the population and 3.2% of families are below the poverty line. Out of the total people living in poverty, 4.3% are under the age of 18 and 4.4% are 65 or older.

Figure 2. Bot-created article compared to a human-edited article. The upper screenshot is the La Grange, Illinois article as created by rambot on 11 December 2002. The lower screenshot shows the same article on 12 January 2009. Available at: http://en.wikipedia.org/wiki/La_Grange,_Illinois

La Grange, Illinois

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by Tom Tallman (talk | contribs) at 15:09, 16 January 2009. It may differ significantly from the current revision.

(diff) ← Previous revision | Current revision (diff) | Newer revision → (diff)

La Grange, a suburb of Chicago, is a village in Cook County, in the U.S. state of Illinois. The population was 15,608 at the 2000 census. The name *La Grange* is French for "the barn."

Contents [hide]
1 History
2 Geography
3 Demographics
4 Government
5 Business and commerce
6 Transportation
7 Architecture
8 Public Schools
9 Health
10 Notable people
11 References
12 External links

History

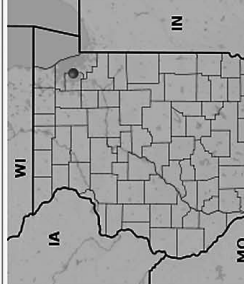
The area around La Grange was first settled in the 1830s, when Chicago residents, already fed up with the rapid population increase in that city in the decade since its incorporation, moved out to the west. The first settler, Robert Leitch, came to what is now La Grange in 1830, a full seven years before the city of Chicago was incorporated in March 1837. La Grange's location, at approximately thirteen miles from Chicago's Loop, is not considered far at all from the city by today's standards, but in that time the residents enjoyed the peace of rural life without much communication with urban residents.

incorporated on June 11, 1879, the Village of La Grange was the dream of Franklin Dwight Cossitt, born in Granby, Connecticut and raised in Tennessee, who moved to Chicago in 1892 and built a successful wholesale grocery business.

In 1870, Cossitt purchased several hundred acres of farmland in Lyons Township, along the Chicago-Dixon Road, known today as Ogden Avenue (U.S. Highway 354). Ogden Avenue, on the site of a defunct Native American trail, was also referred to as the "Old Plank Road". Planks were often stolen by settlers to be used as building material, which made traveling very bumpy. When the Chicago, Burlington & Quincy Railroad came to town, La Grange was a milk stop called Hazel Glen. A few miles to the south, through present day Willow Springs, the Illinois and Michigan Canal had emerged as a major shipping corridor, connecting Chicago and the Great Lakes with the Illinois and Mississippi Rivers.

Cossitt set out to build the ideal suburban village - laying out streets, planting trees, donating property for churches and schools, and building quality homes for sale. He also placed liquor restrictions in the land deeds he sold to prevent the village from becoming a saloon town.

When Cossitt began his development, the area was served by a post office known as Kensington. But upon learning of another community already with that name in Illinois, Cossitt decided to name his town in honor of La Grange, Tennessee, where he had been raised as a youth on an uncle's cotton farm. However, today Kensington remains the name of one of the village's major streets.

La Grange Village			
La Grange Village Hall			
Country	United States		
State	Illinois		
County	Cook		
Coordinates	 41°48'29"N 87°52'24"W		
Area - land	2.5 sq mi (6 km ²)		
Density	2.5 sq mi (6 km ²)		
Population	15,908 (2000)		
Density	6,220.7/sq mi (2,402 /km ²)		
Timezone	GST (UTC-6)		
- summer (DST)	CDT (UTC-5)		
Postal code	60525		
Area code	708		
			

Wikipedia Bot Activity in Most-Used Languages Worldwide

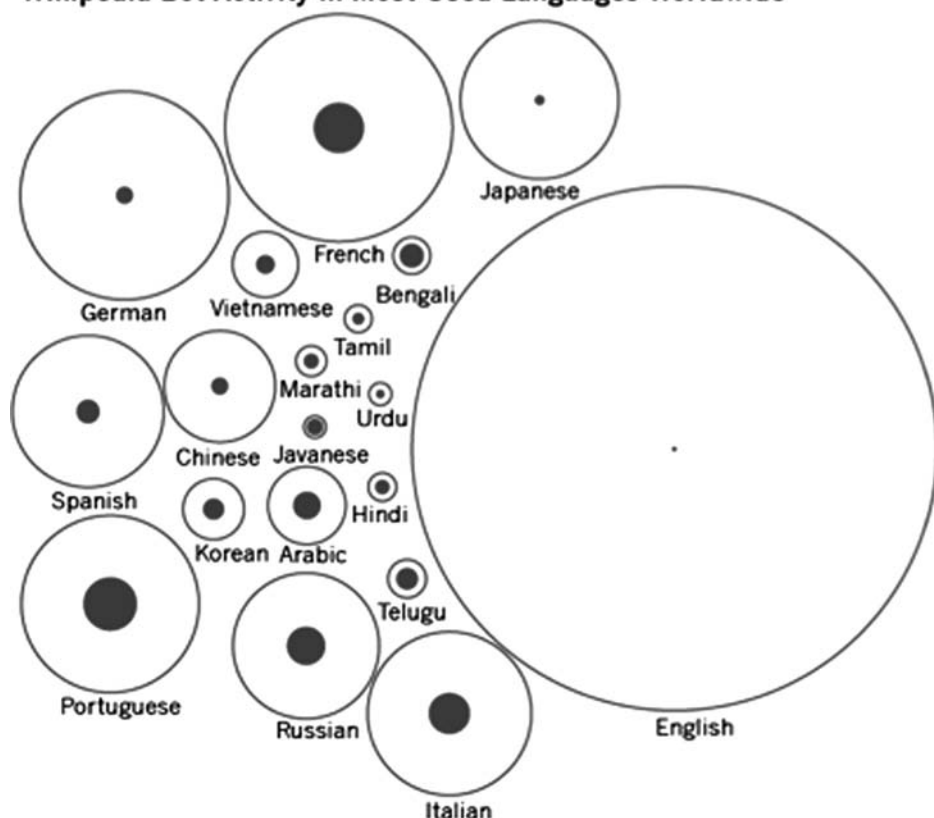


Figure 3. Visualization of Wikipedia bot activity in most-used languages worldwide, overview and detail. The outer circle is the Wikipedia size in that specific language; the inner dot depicts the percentage of bot activity in that language Wikipedia.

Source: Rogers et al. (2008). Graphic by Auke Touwslager using the Dorling Map Tool, Digital Methods Initiative, Amsterdam, 2009. Available at: <http://wiki.digitalmethods.net/Dmi/NetworkedContent>

version even less. Wikipedias of small and endangered languages show a high dependency on bots and a relatively small percentage of human edits. Oriya, for instance, depends 89 per cent on automated software programmes; one small Wikipedia, in the language Bishnupriya Manipuri has seen 97 per cent of its edits made by bots (<http://stats.wikimedia.org/EN/TablesWikipediaBPY.htm>).

Further analysis of bot activity versus human activity reveals that the variety of bot dependency can be an indicator of the state of the language Wikipedia, if not the state of that language, in the global constellation. Looking at the types of bots, we may notice that Wikipedias are maintained mainly by bots that network the content, so-called interwiki and interlanguage bots. Phrased differently, the bots active in these spaces take care of linking articles to articles in Wikipedias, to prevent them from becoming 'orphans' or

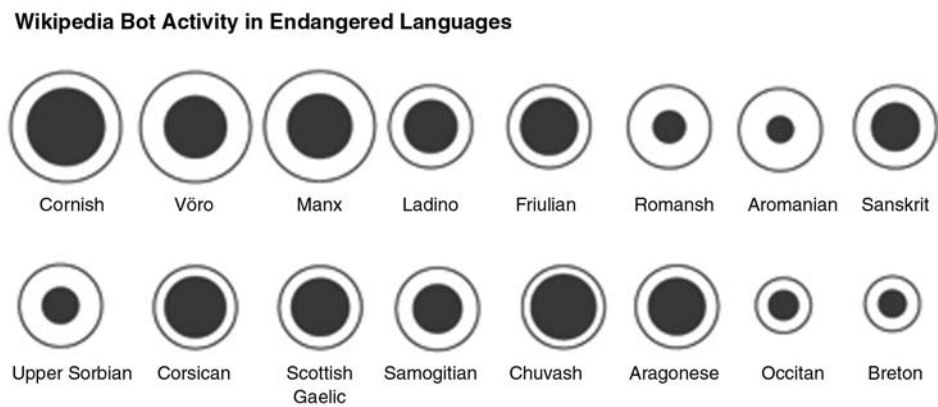


Figure 4. Bot activity in endangered language Wikipedias. Analysis: Rogers et al. (2008). Graphic by Auke Touwslager using the Dorling Map Tool, Digital Methods Initiative, Amsterdam, 2009. Available at: <http://wiki.digitalmethods.net/Dmi/NetworkedContent>

dead ends. Wikipedia policy states that articles should be networked and be part of the Wikipedia web. This core principle is summarized as: ‘Link articles sideways to neighbors, upwards to categories and contexts, and downwards to sub-articles to create a useful web of information’ (http://en.wikipedia.org/wiki/Wikipedia:Build_the_web). Not only are ‘good’ Wikipedia articles full of links to reliable sources, but they should also link to related Wikipedia articles and sub-articles and *be linked to* themselves. Articles that only refer to each other, but are not linked to or linking to other articles, are also considered a threat to the principle of building the web.

We can analyze a language’s state of interconnectedness using the Wikipedia statistics pages, featuring lists of the most active bots per language Wikipedia. They reveal that most-used language Wikipedias, which obviously contain much more content than the smaller language Wikipedias, have bot activity distributed across administrative tasks. In German, for instance, the top 45 listing of most active bots features 27 interwiki bots and 18 bots that are meant to edit content, add categories and fix broken links.⁸ In the smaller language Wikipedias, bots significantly outnumber human editors and are mostly dedicated to linking articles to related articles in other Wikipedias. They make sure that the content, however scarce, is networked. The Cornish Wikipedia’s top 45 of most active bots, for instance, shows at least 35 interwiki bots, while the remainder are bots with unspecified functions. These interwiki bots, such as ‘Silvonenbot’, a bot that adds inter-language links, make connections between various language Wikipedias. Smaller language Wikipedias thus make sure that every article is properly linked sideways and prevent the language Wikipedia from becoming isolated.

Tracing the collaboration between human and non-human agents in Wikipedias thus allows for an interesting and unexpected insight into the culturally and linguistically diverse makeup of this global project. Following the ‘wisdom of crowds’ paradigm, we might have been tempted to look for cultural-linguistic diversity in terms of many people across the world collaborating in different languages and from a number of cultural backgrounds. In line with this paradigm, British information scientists have demonstrated that

the internet (and Wikipedia, in particular) is anything but a culturally neutral space; major aspects of collaborative online work are influenced by pre-existing cultural differences between human contributors (Pfeil et al., 2006). Adding a 'natively digital' analysis of the varied distribution of bot dependency across the wide range of language Wikipedias, we show that cultural differences in collaborative authoring of Wikipedia content cannot just be accounted for in terms of its *human* users; they reveal themselves perhaps more candidly in the relative shares of human and non-human contributions as revealed through automated patterns of contributions. High levels of bot activity, mainly dedicated to networking content and to building the web, are an indicator of small or endangered languages. A richer variety of bot activity, largely subservient to human edit activity, could be considered an indicator of a large and lively language space.

Conclusion

Wikipedia has most commonly been evaluated – either praised or criticized – for its collaborative knowledge production by many (anonymous) minds. From the start, researchers have pointed out that Wikipedia thrived by virtue of a small core of dedicated contributors rather than a large crowd of collaborators, even if the encyclopaedia became more hospitable to common users after 2006. In addition, Wikipedia has never been this mythical egalitarian space, for the various user groups have very distinct levels of permissions. Past Wikipedia research has focused mainly on the crowdsourcing of knowledge as well as on the reliability of Wikipedia content. As we have shown, some researchers, such as Halavais and Magnus, tested Wikipedia by entering false information. These types of research have isolated Wikipedia content as a static product mainly by assessing it against other encyclopaedic records.⁹ Other research, mostly (investigative) journalism, was intent on 'outing' anonymous human editors, making use of 'counter-technology' like the WikiScanner to reveal the identity of contributors or the origin of an edit. More recently, we have seen different research approaches to Wikipedia, such as that of Rosenzweig who acknowledges the encyclopaedia's dynamic content as well as the significance of its partially automated content-management system. For the most part, though, scholarly evaluations of Wikipedia have adhered to the human content-agent paradigm.

In this article, we have argued that Wikipedia's nature and quality should be evaluated in terms of collaborative qualities, not only of its human users, but specifically of its human *and* non-human actors. Since 2002, Wikipedia content has been maintained by both tool-assisted human editors and bots, and collaboration has been modulated by protocols and strict managerial hierarchies. Bots are systematically deployed to detect and revert vandalism, monitor certain articles and, if necessary, ban users, but they also play a substantial role in the creation and maintenance of entries. As we have shown, bot activity may be analyzed as an indicator of the international or intercultural dimension of Wikipedia as a global project.

In the fall of 2009, Wikipedia has introduced WikiTrust, a MediaWiki extension developed by the WikiLab at the University of California in Santa Cruz. With WikiTrust, newly edited parts of Wikipedia articles are colour coded according to reliability, based on the author's reputation, which is established by the lifespan of their other contributions.

Instead of turning to the expert to check all articles, Wikipedia builds further on the combination of rules, hierarchies and editors. To understand Wikipedia's collaborative process, we need to unravel not simply Wikipedia's human agents, but the specificities of its technicity.

We propose to extend this kind of analysis from Wikipedia to various kinds of Web 2.0 infrastructures. Non-human actors and coded protocols are often overlooked in the many optimistic Web 2.0 theories that triumphantly claim the virtues of mass collaboration on Web 2.0 platforms (Tapscott and Williams, 2006). It is important to question the assumptions of the internet as a merely social laboratory of human interaction, instead analyzing and interrogating the sociotechnical system that lies at the core of Web 2.0 platforms. Human and machine contributions are complementary parts of a society of control in which social interactions are increasingly facilitated by means of coded, automated processes. Human judgements such as reliability, accuracy or factuality are turned into machine-coded and regulated alert systems, as illustrated by Wikipedia's mechanisms for generating and checking content. Nicolas Carr compares Web 2.0 to a Mechanical Turk, which 'turns people's actions and judgments into functions in a software program' (2008: 218). A thorough and critical understanding of the automated processes that structure human judgements and decisions requires analytical skills and medium-specific methods which are crucial to a full understanding of how the internet works. Instead of succumbing to the mechanisms of control, users should learn to critically analyse their interactions with technology and actively engage in technology's development (Zittrain, 2008: 245).

In line with David Beer's call in this journal for a more thorough understanding of the 'technological unconsciousness' of participatory web cultures, we have tried to explain the 'performative infrastructure' of Wikipedia by focusing on the second level of analysis that Beer proposes; namely, unravelling software infrastructures and their applications (Beer, 2009: 998). In this article, we have deployed several natively digital methods to unravel in detail the close interdependency of human and technological agents. It is important to comprehend the powerful information technologies that shape our everyday life and the coded mechanisms behind our informational practices and cultural experiences. The analysis of the Wikipedia platform as a sociotechnical system is a first step in that direction.

Notes

- 1 'Digital methods' is a term for medium-specific methods for web research coined by Richard Rogers (2009). The research for this article was conducted with the Digital Methods Initiative (2008) and the Govcom.org Foundation (2008).
- 2 On top of the set of rules, there is a fourth important general principle emphasizing the ideals of openness and collaboration that lie at the core of the project: ignore all rules. This general principle was written up by Larry Sanger to make clear that, above all, Wikipedia is an open platform. Wikipedians should first and foremost strive to improve and maintain Wikipedia (see: http://en.wikipedia.org/wiki/Wikipedia:Ignore_all_rules).
- 3 Halavais's approach was heavily criticized, mainly because he deliberately littered his object of study. After the event, Halavais regretted his approach, especially because the media attention on

his experiment encouraged others to test Wikipedia by inserting mistakes. His website (<http://alex.halavais.net/the-isuzu-experiment>) now includes a call for testing Wikipedia in a 'non-destructive way'.

- 4 In the summer of 2008, Virgil Griffith launched the WikiWatcher suite, a set of tools designed for monitoring and maintaining Wikipedia. The suite includes a tool that makes it possible to de-anonymize users with a username whose IP addresses match those of other user(name)s or companies/institutions in an IP-to-Geo database. This stretches the notion of anonymity from the unregistered to the registered with a username (see Griffith, 2008b).
- 5 WikiScanner 2 works the other way around: enter a company name or URL and the tool shows you which Wikipedia articles were edited from that organization's IP address.
- 6 An overview of the available local Wikipedias is given on the Wikipedia portal page (www.wikipedia.org). Wikipedia currently has 264 language versions, some of which only have a main page and no articles as of yet. The largest Wikipedia is in English, with more than 2 million articles; followed by the German, French, Polish and Japanese editions, each of which contain more than half a million articles. Seventeen other language editions contain 100,000+ articles, and more than 100 other languages contain 1000+ articles; the overview also includes the smaller ones with only 100+ articles and even Wikipedias that have only a main page (http://wikimediafoundation.org/wiki/Our_projects).
- 7 In a case study on Wikipedia as 'networked content' that Sabine Niederer conducted with Richard Rogers et al. (2008), during the 10-Year Jubilee Workshop of the Govcom.org Foundation, the researchers noticed the great discrepancy between bot activity in English and the other language versions of Wikipedia.
- 8 The top 45 most active bots on the German Wikipedia consist of 27 interwiki bots and 18 bots with various tasks such as editing, fixing links and adding categories (<http://stats.wikimedia.org/EN/TablesWikipediaDE.htm>).
- 9 To this date, communication scholars like Halavais and Lackaff (2008) examine Wikipedia's reliability and completeness, assessing the qualities of its users rather than those of its systems.

References

- Adler T, de Alfaro L, Pye I, and Raman W (2008) Measuring author contributions to Wikipedia. In: *Proceedings of WikiSym 2008*, Porto, 8–10 September. New York: ACM. Available at: <http://users.soe.ucsc.edu/~luca/papers/08/wikisym08-users.pdf>
- Baker N (2008) The charms of Wikipedia. *New York Review of Books* 55(4). Available at: <http://www.nybooks.com/articles/21131>
- BBC (2005) Wikipedia survives research test. *BBC News* (15 December). Available at: <http://news.bbc.co.uk/2/hi/technology/4530930.stm>
- Beer D (2009) Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11(6): 985–1002.
- Benkler Y (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press.
- boyd d (2005) Academia and Wikipedia. *Corante* (4 January). Available at: http://many.corante.com/archives/2005/01/04/academia_and_wikipedia.php
- Bruns A (2008) *Blogs, Wikipedia, Second Life, and Beyond: From Production to Prodisusage*. New York: Peter Lang. Available at: <http://produsage.org/>

- Burke M, Kraut R (2008) Taking up the mop: Identifying future Wikipedia administrators. In: *Proceedings of the 2008 CHI Conference*, Florence, 5–10 April. New York: ACM, 3441–6. Available at: <http://portal.acm.org/citation.cfm?id=1358628.1358871>
- Carr T (2008) *The Big Switch: Rewiring the World, from Edison to Google*. New York: Norton.
- Chesney T (2006) An empirical examination of Wikipedia's credibility. *First Monday* 11(11). Available at: http://firstmonday.org/issues/issue11_11/chesney/
- Deleuze G (1990) Society of control. *L'autre journal* (1). Available at: <http://www.nadir.org/nadir/archiv/netzkritik/societyofcontrol.html>
- Denning P, Horning J, Parnas D, and Weinstein L (2005) Inside risks: Wikipedia risks. *Communications of the ACM* 48(12): 152.
- Digital Methods Initiative (2008) *Digital Methods Initiative*. Available at: <http://www.digital-methods.net>
- Galloway A (2004) *Protocol: How Control Exists after Decentralization*. Cambridge, MA: MIT Press.
- Ghosh RA, Prakash VV (2000) Orbiten free software survey. *First Monday* 5(7). Available at: http://www.firstmonday.org/issues/issue5_7/ghosh/
- Giles J (2005) Internet encyclopaedias go head to head. *Nature* 438: 900–1. Available at: <http://www.nature.com/nature/journal/v438/n7070/full/438900a.html>
- Govcom.org Foundation (2008) Govcom.org 10 year jubilee workshop. *Digital Methods Initiative* (11–15 August). Available at: http://wiki.digitalmethods.net/Dmi/GovcomorgJubilee#Jubilee_Programme_archivedcx
- Griffith V (2008a) WikiScanner homepage. Available at: <http://virgil.gr/31.html>
- Griffith V (2008b) Wikiwatcher homepage. Available at: <http://wikiwatcher.virgil.gr/>
- Halavais A (2004) The Isuzu experiment. In: *A Thaumaturgical Compendium Blog*. Available at: <http://alex.halavais.net/the-isuzu-experiment/>
- Halavais A, Lackaff D (2008) An analysis of topical coverage of Wikipedia. *Journal of Computer-mediated Communication* 13: 429–40.
- Howe J (2006) The rise of crowdsourcing. *Wired* 14(6). Available at: <http://www.wired.com/wired/archive/14.06/crowds.html>
- Jenkins H (2006) *Convergence Culture: Where Old and New Media Collide*. Cambridge, MA: MIT Press.
- Keen A (2008) *The Cult of the Amateur: How Blogs, MySpace, YouTube, and the Rest of Today's User-generated Media Are Killing Our Culture and Economy*. London: Nicholas Brealey.
- Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in Wikipedia: Quality through co-ordination. In: *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work*. New York: ACM, 37–46.
- Kittur A, Chi E, Pendleton B, Sun B, and Mytkowicz T (2008) Power of the few vs wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. Paper presented at 'CHI 2007', San Jose, 28 April–3 May.
- Latour B (1991) Technology is society made durable. In: Law J (ed.) *A Sociology of Monsters: Essays on Power, Technology and Domination*. London: Routledge, 103–32.
- Magnus PD (2008) Early response to false claims in Wikipedia. *First Monday* 13(9). Available at: <http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2115/2027>
- Negroponte N (1995) *Being Digital*. New York: Vintage.

- Pfeil U, Zaphiris P, and Ang CS (2006) 'Cultural differences in collaborative authoring of Wikipedia', *Journal of Computer-mediated Communication* 12: 88–113.
- Poe M (2006) The hive. *Atlantic Online* (September). Available at: <http://www.theatlantic.com/doc/200609/wikipedia>
- Rogers R (2009) *The End of the Virtual: Digital Methods*. Amsterdam: Vossiuspers UvA.
- Rogers R, Niederer S, Deveraux Z, and Nijhof B (2008) Networked content. *Digital Methods Initiative*. Available at: <http://wiki.digitalmethods.net/Dmi/NetworkedContent>
- Rosenzweig R (2006) Can history be open source? Wikipedia and the future of the past. *Journal of American History* 93(1): 117–46.
- Sirky C (2005) K5 article on Wikipedia anti-elitism. *Corante*. Available at: http://many.corante.com/archives/2005/01/03/k5_article_on_wikipedia_antielitism.php
- Shirky C (2008) *Here Comes Everybody: The Power of Organizing without Organizations*. New York: Penguin.
- Stalder F, Hirsch J (2002) Open source intelligence. *First Monday* 7(6–3). Available at: <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/961>
- Stevenson M, Rogers R (2009) Digital methods: First steps. *East Review* 28(2). Available at: <http://www.easst.net/review/june2009/stevenson>
- Sunstein CR (2006) *Infotopia: How Many Minds Produce Knowledge*. Oxford: Oxford University Press.
- Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Societies and Nations*. New York: Doubleday.
- Swartz A (2006) Who writes Wikipedia? *Raw Thought Blog*. Available at: <http://www.aaronsw.com/weblog/whowriteswikipedia/>
- Tapscott D, Williams AD (2006) *Wikinomics: How Mass Collaboration Changes Everything*. New York: Penguin.
- Zittrain J (2008) *The Future of the Internet*. New York: Penguin.

Sabine Niederer is PhD candidate in Media Studies at the University of Amsterdam and member of the Digital Methods Initiative, Amsterdam. She is also Managing Director of the Institute of Network Cultures, the new media research centre at the Amsterdam University of Applied Sciences.

José van Dijck is a Professor of Media and Culture at the University of Amsterdam where she is currently Dean of Humanities. She has published widely in the area of media and science, (digital) media technologies, and television and culture. Her latest book is entitled *Mediated Memories in the Digital Age* (Stanford University Press, 2007). Address: University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, the Netherlands. [email: j.f.t.m.vanDijck@uva.nl]

J.M. Reagle, (2008). In good faith: Wikipedia and the pursuit of the universal encyclopedia, PhD dissertation, New York University, pp. 70-102.

Sponsoring Committee: Professor Helen Nissenbaum, Chairperson
Professor Gabriella Coleman
Professor Natalia Levina

IN GOOD FAITH:
WIKIPEDIA COLLABORATION AND
THE PURSUIT OF THE UNIVERSAL ENCYCLOPEDIA

Joseph M. Reagle Jr.

Department of Media, Culture, and Communication

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in the
Steinhardt School of Culture, Education, and Human Development
New York University
2008

COMMUNITY AND COLLABORATION

In some ways, as previous chapters show, Wikipedia resembles its predecessors. But what, then, makes it different? I answer this question in three parts. First, the community that produces it is unlike that of any other encyclopedic undertaking. I describe the Wikipedia community, and other related projects, as an open content community. Second, I consider the particularities of Wikipedia collaborative culture and the centrality of notions such as “Neutral Point of View” and “good faith.” Finally, I posit a type of leadership within open content communities. I believe these models of community, culture, and leadership capture Wikipedia’s collaborative culture, and delineate some of the similarities, and novelties of Wikipedia relative to other projects. For me, they also give a sense of some of the practical difficulties inherent in the vision of Otlet and Wells, the problems encountered by a global project attempting to increase access to knowledge and further social accord.

Portions of chapter 4, 5, and 6 have appeared as fragments or in earlier form where I’ve addressed questions of openness (Reagle, 2006b,c, 2007b), neutrality (Reagle, 2005a, 2006a), and leadership (Reagle, 2007a).

CHAPTER IV

WIKIPEDIA AS AN OPEN CONTENT COMMUNITY

At the outset I claimed that “Wikipedia is a community and the encyclopedia is a snapshot of its continuing conversation.” But the collection of people involved in Wikipedia is different from that of any other encyclopedic undertaking. So how might we understand this particular community? I believe to understand Wikipedia one needs to appreciate it as an *open content community*. This is a concept I’ve developed elsewhere (Reagle, 2004) so as to distinguish between an open (or free) type of content, namely Free and Open Source Software (FOSS), and the community that produces it—because a “closed” group, such as a company, can produce software under a “open” license. Furthermore, “open” has become a bit of a buzz word, describing everything from democracy (Rushkoff, 2003) to religion (Krangel, 2007)—I review much of this usage in earlier work as well (Reagle, 2006b). Additionally, when contemporary sources speak of openness as an attribute of community, it is often in an inexact or overly simplistic way. Consider the slogan on the welcoming page of the English Wikipedia: “Wikipedia, the free encyclopedia that anyone can edit” (Wikipedia, 2006e). It would appear that the universal vision of increased access to information now includes its production as well as consumption. Yet, open source type projects like the Linux kernel, Apache Web server, and Wikipedia are often characterized incorrectly by way of an inappropriate, if not naïve, extreme. A utopian rendering of openness is that “anything goes”: there are no community structures or norms, anyone can do anything they please.

This understanding of “anything goes” is untenable: some level of struc-

ture is inevitable in social relations, and often necessary to support other values. In his 1911 book *Political Parties*, Rober Michels (2001, 6.2) wrote of the development of an oligarchy within democratic parties as an “Iron Law.” In 1970 Jo Freeman (1996, p. 1) wrote about the “tyranny” present in seemingly egalitarian feminist groups of the earlier decade: “‘Structurelessness’ is organisationally impossible. We cannot decide whether to have a structured or structureless group; only whether or not to have a formally structured one.” And more recently, Mitch Kapor expressed a similar sentiment with respect to the early management of the Internet when he noted that “Inside every working anarchy, there’s an Old Boy Network” (as cited in Reagle, 1999). Yet, this does not mean the notion of openness should be jettisoned altogether. The prevalence of the term “open” in contemporary discourse, arising largely from the popular attention on FOSS, is indicative of something important, as is coming to a better understanding of what it means for community, like Wikipedia, to be open or not.

To address this question of openness and Wikipedia, I first provide a thumbnail sketch of the scope of Wikipedia community. I then consider Wikipedia in light of five criteria I’ve previously specified for an open content community. This is juxtaposed with three cases in which Wikipedia’s openness is challenged. In the first case I return to the question of whether Wikipedia is really something “anyone can edit”? That is, when Wikipedia implemented new technical features to help limit vandalization of the site, did it make Wikipedia more or less open? In the second case I describe the way in which a maturing open content community’s requirement to interact with the sometimes “closed” world of law affects its openness. In this case, I review Wikipedia’s “office action” in which agents of Wikipedia act privately so as to mitigate legal problems though this is contrary to the community values of deliberation and transparency. Finally, I explore case in which a closed (female only) group is set up outside of, and perhaps because of,

the “openness” of the larger Wikipedia community.

The Wikipedia Community

After a couple of decades’ worth of literature on online communities, the proposition that there is a community associated with Wikipedia should be unremarkable. Granted, what exactly constitutes the Wikipedia community is an interesting question. For example, does the person who occasionally corrects a spelling count as a community member? In the previous chapter I do provide some references on the character of Wikipedia contribution, but more generally, for the purposes of this project I use the word community to speak of a group of interdependent people who “participate together in discussion and decision making and who share certain practices that both define the community and are restored by it” (Bellah et al., 1996, p. 333). Wikipedia community members do share common practices and norms, these are part of the collaborative culture discussed in the next chapter. Furthermore, the Wikipedia community can be further understood as “prosocial” (Bowles and Gintis, 1998; Sproull et al., 2004) in that exhibits behavior that is intentional, voluntary, and of benefit to others.

That said, do we have any sense of the scope and scale of the community? The “About” Wikipedia article states “There are more than 75,000 active contributors working on some 8,700,000 articles in more than 250 languages” (Wikipedia, 2007av). For those involved in administrative functions (e.g., protecting pages), “There are 1,382 (as of 8 November 2007) users with sysop rights (active and otherwise), 938 of them active (as of 2007-11-08)” (Wikipedia, 2007bf). In November 2007, I count almost 200 names on the #wikipedia chat channel and there are almost 700 subscribers to the community *Wikizine* bulletin (Walter, 2007). On the wikipedia-l (relatively low traffic relative to wikiEN-l) I count

over 300 unique posters from February 2005 to November 2007, though I am confident this is only a fraction of those subscribed. More topically, “Wikipedia Projects” are wiki pages in which contributors interested in a particular topic can plan and discuss their efforts (Wikipedia, 2007bq), of which there are over 200 on the English Wikipedia (Wikipedia, 2007bg). Furthermore, through Wikipedia “meetups” I’ve attended in New York and the annual Wikimania conferences I’ve met a couple dozen people. Many of these people I’ve spoken to more than once, and it’s quite easy to speak to a newly met Wikipedian about issues of concern to the community. In addition to the email lists and *Wikizine* there are other community fora such as the popular *Wikipedia Signpost* (Wikipedia, 2007bp), and various wiki blogs and their aggregators (Foundation, 2007a; Millosh, 2007; Open, 2007). Plus, there are dozens of communities around the different language Wikipedias. Therefore, I believe within the larger community of tens of thousands of active contributors who are familiar with the basic practices and norms of Wikipedia, there are also smaller communities on the scales of hundreds or dozens of members within language, geographical, functional, and topical boundaries.

In any case, my intention is not paint a demographic portrait of Wikipedia contributors but to provide a sense of the larger community and how it views itself on questions of openness. In “Notions of Openness” (Reagle, 2006c) I review the many uses of “open” as inspired by the success of Free and Open Source Software; from this I discern three shared features that are represented in the characterization of “openness” as an *accessible and flexible* type of *collaboration* whose result may be *widely shared*. I further specify that an open content community (Reagle, 2004) is one that delivers or demonstrates the following:

1. Open products: provides products which are available under licenses like those that satisfy the Open Source Definition.
2. Transparency: makes its processes, rules, determinations, and their ratio-

nales available.

3. Integrity: ensures the integrity of the processes and the participants' contributions.
4. Non-discrimination: prohibits arbitrary discrimination against persons, groups, or characteristics not relevant to the community's scope of activity. Persons and proposals should be judged on their merits. Leadership should be based on meritocratic or representative processes.
5. Non-interference: the linchpin of openness, if a constituency disagrees with the implementation of the previous three criteria, the first criteria permits them to take the products and commence work on them under their own conceptualization without interference. While "forking" is often complained about in open communities—it can create some redundancy/inefficiency—it is an essential characteristic and major benefit of open communities as well.

Although the first criterion provides a "bright line" with which one can distinguish between open products and their licenses, the social criteria of transparency, integrity, and nondiscrimination do not provide for an equally clear demarcation. (What counts as open or free content has not always been an easy question either, but at least we now have the "Free Software Definition" (Stallman, 2005e) and "Open Source Definition" (OSI, 2006b) to rely upon.) Indeed, a common behavior of an open community is the self-reflective discourse of what it means to be open on difficult boundary cases. Consequently, I argue that a test of an open community is if a constituency that is dissatisfied with results of such a discussion can fork (copy and relocate) the work elsewhere. Additionally, although the often voluntary character of the community is not explicitly articulated in my conceptualization, it is important to note that voluntariness is critical to

understanding the moral/ideological light in which many of the members view their participation.

In the following sections I present the Wikipedia community against the criteria of an open content community.

Open Products

What is often meant by the term “open” is a generalization from the FOSS movement. Communities marshaling themselves under these banners cooperatively produce, in public view, software, technical standards, or other content that is intended to be widely shared. Fortunately, there are now a number of excellent scholarly resources on the FOSS phenomenon (see Dibona et al., 1999; Williams, 2002; Weber, 2004; Chopra and Dexter, 2007); because of this I will provide only the briefest description of its history so as to understand what is meant by “open products.”

The Free Software movement was spearheaded by Richard Stallman at MIT in the 1980s. When Stallman found it difficult to obtain the source code of a troublesome Xerox printer, he feared that the freedom to tinker and improve technology were being challenged by a different, proprietary, conceptualization of information (Williams, 2002). To respond to this shift he created two organizations: the GNU Project in 1984, which develops and maintains free software, and the Free Software Foundation (FSF) in 1985 (Stallman, 2005f), which houses legal and advocacy efforts. Perhaps most importantly he wrote the first version of the GNU General Public License (GPL) in 1989. The GPL is the seminal copyright license for “free software”; it ensures that the “freedom” associated with being able to access and modify software is maintained with the original software and its derivations. It has important safeguards, including its famous reciprocal

provision: if you modify and distribute software obtained under the GPL license, your derivation also must be publicly accessible and licensed under the GPL. (Ensuring the “busy bees” of knowledge production referred to in the previous chapter are not hindered in their information cross-pollination.)

In 1991, Linus Torvalds started development of Linux: a UNIX like operating system kernel, the core computer program that mediates between applications and the underlying hardware. While it was not part of the GNU Project, and differed in design philosophy from the GNU’s kernel (named “Hurd”), it was released under the GPL. While Stallman’s stance on “freedom” is more ideological, Torvalds approach is more pragmatic. Furthermore, other projects, such as the Apache web server, and eventually Netscape’s Mozilla web browser, were being developed under similar open licenses except that, unlike the GPL, they often permit proprietary derivations. With such a license, a company may take open source software, change it, and include it in their product without releasing their changes back to the community.

The tension between the ideology of free software and its other, additional, benefits led to the concept of Open Source in 1998. The Open Source Initiative (OSI) was founded when Netscape was considering the release of its browser as free software; participants at the meeting “decided it was time to dump the moralizing and confrontational attitude that had been associated with ‘free software’ in the past and sell the idea strictly on the same pragmatic, business-case grounds that had motivated Netscape. They brainstormed about tactics and a new label. ‘Open source’, contributed by Chris Peterson, was the best thing they came up with” (OSI, 2006a). Under the Open Source banner the language and ideology of freedom was sidelined so as to highlight pragmatic benefits and increase corporate involvement.

The benefits of openness are not limited to software. Because the doc-

umentation that accompanies free software should also be free the FSF created the GNU Free Documentation License (GFDL) in 1999. Of course, in the new millennium this model of openness has extended to forms of cultural production beyond technical content. Wikipedia's co-founder, Jimmy Wales, has stated that a seminal article by Eric Raymond (1997) that likened FOSS production to that of a vibrant decentralized bazaar "opened my eyes to the possibility of mass collaboration" (as cited in Schiff, 2006, p. 3). In fact, in October 2001, not even a full year old, Wales collected those principles he thought were responsible, and would continue to be needed, for its success. In his "Statement of Principles," Wales (2001c) wrote that "Wikipedia's success to date is entirely a function of our open community." As Nupedia and Wikipedia were licensed under the GFDL from the start, "The GNU FDL license, the openness and viral nature of it, are fundamental to the long-term success of the site."

Ironically, Wikipedia is not looked upon favorably by some prominent FOSS developers. Eric Raymond himself has characterized Wikipedia as a "disaster" that is "infested with moonbats" (as cited in Schiff, 2006, p. 8); in this view Wikipedia is an unsuitable case of the open source model because the merit of software developers and their code can be judged by objective standards (e.g., speed or efficiency), but knowledge claims can not. A participant on the geek discussion site Kuro5hin writes, "People love to compare Wikipedia to Open Source but guess what: bad, incorrect code doesn't compile. Bad, incorrect information on the 'net lives on and non-experts hardly ever notice the mistake" (dharma, 2004). This difference between functional and expressive content is one of the many possible differences between "open source" and "open culture" as Felix Stalder (2006) puts it.

In time, because the GFDL was intended to accompany the textual documentation of software, and was perceived by some as not being flexible

enough, new non-software content licenses have appeared. More widely, the Creative Commons project, launched in 2001, provides licenses for the sharing of texts, photos, and music. Lawrence Lessig (2004), a founder of Creative Commons, helped popularize the notion of freedom and openness in domains beyond software with his book *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. Wikipedia is probably the best known “open product” of the wider free culture movement today.

Transparency and Integrity

“An open content community should make its processes, rules, determinations, and their rationales available; the integrity of those processes and the participants’ contributions should be respected.” At first blush, and for the vast majority of cases, communities based on public wikis should do well on the criteria of transparency and integrity, particularly when you compare them to traditional organizations. On the Meatball wiki, “a common space for wiki developers and proprietors from all over the Internet to collaborate” (Meatball, 2006b), this criteria is partially captured by what it calls “Fair Process” (Meatball, 2007a) which itself includes the three principles of engagement, explanation, and clarity; fair process is particularly important in voluntary communities “because fair process builds trust and commitment, people will go above and beyond the call of duty” (Meatball, 2007a).

While some warn of “eroding accountability in computerized societies” (Nissenbaum, 1996), others have argued that the open development of FOSS may be an exception, and even provide a model for achieving accountability for other technologies or institutions (David, 2004). Consequently it shouldn’t be surprising that transparency has come to be an attributed feature of Wikipedia.

Jill Coffin explains that transparency “allows participants to understand the reasoning behind decisions, contributing to trust in the Wikipedia process. It also allows newbies a means to understand informal community protocol and culture, as well as reduce abusive practice” (Coffin, 2006). Wiki technology and culture promote the documentation of proposals, discussions, and decisions—everything actually. Integrity can then flow from the accountability inherent to such transparency: the record is there for all to see. Coffin relates this to a famous Linux aphorism: “Schlock and chaos are avoided due to the watchful eyes of the many, exemplifying Linus’ Law, coined and articulated by hacker Eric Raymond as ‘Given enough eyes, all bugs are shallow’” (Coffin, 2006). The importance and hoped for effects of transparency can be seen in the expectations of Wikipedia Stewards, who have significant power in administering all other user rights; they are expected to act transparently:

Steward activity is visible in the Meta rights log. When a request is fulfilled, stewards should note what they did at the local request page (each new request should be accompanied by a link to this) or on the Meta request page.

Steward discussions should occur on Meta, rather than by email, so people can understand the stewards’ decisions and ways of working. (Wikimedia, 2007b)

However, just as a naïve rendering of openness as “anything goes” is overly simplistic, so is the sense that just because something has been posted on the web then one has achieved a perfect level of accountability.

Non-Discrimination

“The criterion of non-discrimination prohibits arbitrary discrimination against persons, groups, or characteristics not relevant to the community’s scope of activity. Persons and proposals should be judged on their merits. Leadership

should be based on meritocratic or representative processes.” However, a common tendency in groups is to adopt an in-group/out-group mentality; Wikipedia cultural norms attempt to counter this. In the 2001 “Statement of Principles” Wales wrote “Newcomers are always to be welcomed. There must be no cabal, there must be no elites, there must be no hierarchy or structure which gets in the way of this openness to newcomers” (Wales, 2001c). This is further reflected in the famous Wikipedia maxim “Please Do Not Bite the Newcomers” (Wikipedia, 2006ab).

Beyond newcomers, there are also norms of nondiscrimination with respect to behavior and beliefs. In the *wikien-l* thread entitled “Wikipedia and autism” Tony Sidaway wrote of the treatment of two admittedly difficult contributors: “Both of them have expressed a strong wish to produce work for Wikipedia. Both of them produce articles that appear weird to non-autists. In my opinion, neither represents a threat to Wikipedia commensurate to the treatment they have received” (Sidaway 2005wa). Wikipedians then discussed how they might best work with and encourage such contributors. Also, as seen in the introductory scenario of chapter 1, Wikipedia administrator MattCrypto unblocked a “racialist” because he thought it was unfair to block someone because of their affiliation rather than Wikipedia actions. Even those with problems or criticisms of Wikipedia should be welcomed if they connect in a constructive way:

Anyone with a complaint should be treated with the utmost respect and dignity. They should be encouraged constantly to present their problems in a constructive way in the open forum of the mailing list. Anyone who just complains without foundation, refusing to join the discussion, I am afraid I must simply reject and ignore. Consensus is a partnership between interested parties working positively for a common goal. I must not let the “squeaky wheel” be greased just for being a jerk. (Wales, 2001c)

However, it is interesting to note that the “Statement of Principles” (Wales, 2001c) of October 27, which I think is a seminal articulation of the Wikipedia

ethos, appeared after two other messages relevant to Wikipedia openness. Jimmy Wales and Larry Sanger are often inappropriately placed at extremes of the “crowds versus experts” continuum; however, Sanger has welcomed mass participation under the guidance of experts and Wales has recognized the challenge of mass participation as Wikipedia continued to grow:

One of the wonderful things about the wiki software, and something that has served us very well so far, is that it is totally wide open. I suspect that any significant deviation from that would kill the magic of the process.

On the other hand, we really are moving into uncharted territory. Wikipedia is already, as far as I know, the most active and heavily trafficked wiki to ever exist. It seems a virtual certainty that the wide open model will start to show some strain (primarily from vandalism) as we move forward.

I have this idea that there should be in the software some concept of “old timer” or “karma points”. This would empower some shadowy mysterious elite group of us to do things that might not be possible for newbies. Editing the homepage for example. We already had one instance of very ugly graffiti posted there (a pornographic cartoon). Some principles that we should use if/when we move in that direction. . . . (Wales, 2001b)

This message is a bit of a faux pas on Wales’ part. In a subsequent chapter on leadership I note that an open content community is often led by a “benevolent dictator” and it deals with the anxiety arising from the tension between the egalitarian ethos and autocratic leadership by way of irony and humor. In this message Wales speaks of being a dictator and of a cabal in much the same way—without appreciating the joke doesn’t work when he tells it. One week later Wales was forced to explain:

In a letter to wikipedia-l, I injudiciously used the word ‘cabal’ and made reference to a ‘shadowy mysterious elite’. This was a very poor choice of words on my part. I thought that many or most people would understand it for what it was—the notion of a non-existent cabal, allegedly controlling things, when in fact there is not one, would be well understood.

Let me be clear. In wikipedia, there should be no elites.

All legitimate participants, no matter how much they may disagree on political, philosophical, or other issues, should always be able to edit pages in the same fashion as they can now. Only behavior that truly and clearly rises to the level of vandalism should be fought with extremely cautious uses of software security measures. (Wales, 2001d)

And the following day Wales posted his “Statement of Principles” on the wiki further highlighting the importance of openness to Wikipedia’s success. Even so, fears of a cabal continue to arise every so often; it is human nature and a social inevitability for practice to sometimes fall short of principle: Wikipedians frequently raise concerns about transparency, integrity, and discrimination (see Ronline, 2006; Martin, 2007a). Despite this, relative to the daily tasks of editing and most other practices, I believe Wikipedia satisfies the criteria—though the community’s discourse about these concerns will continue forevermore.

Non-Interference

Simply, if the content is available under an open/free license, those dissatisfied with any of the other criteria, or other issues, can take it and begin work on it within their own community and culture (Meatball, 2007c). Steven Weber (2004) notes the importance of forking by claiming: “The core freedom in free software is precisely and explicitly the right to fork” (p. 159). While I don’t consider it to be the “core freedom” but a critical social *implication* of free content, I do agree it is a “fundamental characteristic” of FOSS and that “to explain the open source process is, in large part, to explain why that [forking] does not happen very often and why it does when it does, as well as what that means for cooperation” (Weber, 2004, p.92). To this end, David Wheeler likens forking to “the ability to call for a vote of no confidence or a labor strike is important. Fundamentally, the ability to create a fork forces project leaders to pay attention to

their constituencies” (Wheeler, 2005).

Forks of Wikipedia content have happened and are even more frequently threatened and discussed. For example, because of a misunderstanding about the possibility of Wikipedia carrying advertising, the Spanish Wikipedia was forked into *Enciclopedia Libre Universal* (Suarez and Ruiz, 2005). (The misunderstanding has since become resolved and Spanish Wikipedia has superseded the fork.) Or, Larry Sanger, dissatisfied with a lack of respect for expert contributors at Wikipedia, has created a new project called Citizendium (Sanger, 2006b) which considered adopting Wikipedia content that it likes and wants to improve. However, definitively settling upon the license Citizendium would use was not a quick or easy process. One concern among some Citizendium contributors was that if they were to use the GFDL license, and therefore able to use (and improve upon) Wikipedia content, Wikipedia could import the improved Citizendium content back into itself. This was unacceptable to those who wished to distinguish themselves and the superiority of their approach. Therefore, as some Wikipedia content had already been adopted, depending on whether Citizendium chose a license compatible with GFDL, Citizendium experimented with the possibility of “unforking” their borrowed content (Sanger, 2007a)—rewriting it from scratch. In December of 2007 Citizendium chose the “Creative Commons Attribution-ShareAlike 3.0 Unported License (CC-by-sa)” which, after an orthogonal effort to make the GFDL compatible with this Creative Commons license, means “Wikipedia and the Citizendium will be able to exchange content easily” (Sanger, 2007b).

Challenges to Wikipedia's Openness

So far I have described the criteria of what I call an open content community and Wikipedia's applicability. In the following sections I focus more carefully on three particular cases which challenge Wikipedia's claims of openness. In the discipline of computer science, one's constructs are most severely tested by "boundary cases," and I think that considering the cases of whether anyone can really edit, office actions, and the WikiChix enclave give insight into what is meant by openness, and Wikipedia's claims upon it.

Can Anyone Really Edit?

As noted, the English Wikipedia declares itself as "the free encyclopedia that anyone can edit" (Wikipedia, 2006e). Presently, this includes the "anonymous," those who don't log in before editing. Despite the common retort that Wikipedia is "not an experiment in anarchy" (Wikipedia, 2006ag), among other things, the feature of openness and anonymous editing continues to be a valued part of Wikipedia's identity: even those who always log in to edit might still support the ability of others to edit without doing so.

Before proceeding with a discussion of how anonymous editing and blocking are employed in discussions about openness, a brief explanation is in order. Every edit to Wikipedia is logged and can be reviewed on the article's history page. Wikipedia contributors may choose to create an account with a name/identity of their choosing: it might be personally identifiable, or a pseudonym. Editors who have not logged in to such an account are often referred to as "anonymous." In the history log, the edit of an anonymous user is attributed to an IP address, the number associated with a user's computer by their Internet service provider. The reason the term anonymous is not strictly correct is that

there have been cases in which these numbers have been used to trace back an edit to a particular computer. For example, the offices of US Congressional representatives have been identified in removing true but embarrassing information about representatives (Lehmann, 2006; Wikipedia, 2007ac). In fact, those who wish to protect their privacy would be better off creating a pseudonym under which they edit. Then, only the few “checkuser” (Wikipedia, 2007ag) Wikipedians who have access to the server logs would be able to determine the IP address of the originating computer.¹

Unfortunately, Wikipedia is continually vandalized. However, there are various automated tools (“bots”) and groups of users (e.g., the “RC Patrol” who keep an eye on recent changes) that roll back or “revert” articles to their previous state. When it becomes clear that a particular user is persistently abusive, administrators may suspend her account for a period, or in serious cases they might ban her altogether (Wikipedia, 2007bb). However, there is little to prevent such a user from creating another account or editing anonymously. Consequently, administrators have the ability to block users based on their IP number. Whereas any given particular block might be contested, it is difficult to conceive of Wikipedia working at all without such a feature. As the “Wikipedia Is Not an Experiment in Anarchy” article states:

Wikipedia is free and open, but restricts both freedom and openness where they interfere with the purpose of creating an encyclopedia. Accordingly, Wikipedia is not a forum for unregulated free speech. The fact that Wikipedia is an open, self-governing project does not mean that any part of its purpose is to explore the viability of anarchistic communities. Our purpose is to build an encyclopedia, not to test the limits of anarchism. (Wikipedia, 2006ag)

¹Because the term “anonymous” is persistently used by the community, despite my caveat that it really means “not logged in,” I will hereafter use it myself without further qualification unless required by the context.

However, since IP numbers sometimes change, or many users may share an IP number, this approach sometimes blocks the innocent; a balance must be struck. This raises the question of “How Computer Systems Embody Values” (Nissenbaum, 2001)? Specifically, to what extent do technical features such as blocking vandals or requiring registration promote or constrain community values such as openness?

Consider the infamous Seigenthaler case of 2005 in which the biographical article of one of Robert Kennedy’s administrative assistants contained the unfounded claim that he was implicated in the assassinations of the Kennedy brothers. Much to the embarrassment of many Wikipedians, Seigenthaler objected in a widely discussed editorial opinion in *USA Today* (Seigenthaler, 2005). After the identity of the “anonymous” contributor was revealed as the author of a “prank gone wrong,” the press reported that Seigenthaler was not holding a grudge, nor supporting a regulatory crackdown on the Internet, but he did fear “Wikipedia is inviting it by its allowing irresponsible vandals to write anything they want about anybody” (Press, 2005).

In a message to one of the Wikipedia lists, Jimmy Wales objected to this as a mischaracterization of Wikipedia, and its openness. Wales argued that to equate openness with defamation is like equating a restaurant’s steak knives with stabbings. To force everyone in the restaurant to be isolated in steel cages because of the possibility of a stabbing would violate the values of “human kindness, benevolence, and a positive sense of community” and, consequently, “I do not accept the spin that Wikipedia ‘allows anyone to write anything’ just because we do not metaphysically prevent it by putting authors in cages” (Wales, 2005k). The question here seems to be to what extent does the phrase “allow anyone to edit” include the possibility of “allowing vandals to defame”? Seigenthaler’s position is that the first does include the latter, whereas Wales seems to be arguing such

a conclusion is misleading with respect to the community's intention and the balance of consequences. I believe that while vandalism is not condoned, and it is even actively repulsed, implementing technical or social structures that would make vandalism *impossible* would conflict with other community values. How might Wikipedia decrease the possibility of vandalism without unduly affecting other values such as openness?

In June 2006, approximately 6 months after the Seigenthaler incident, a new mechanism was deployed: "semi-protection" blocking (Wikipedia, 2006o). Previously, any user, anonymous or logged in, would be prohibited from making contributions if they were doing so from a blocked IP number which had been the source of problems. Now, logged in users associated with the blocked IP numbers would be able to edit (Wikipedia, 2006p). During the almost eight months of discussion about the feature, some had expressed a concern that it was contrary to the value of openness:

Personally, I think the new blocking policy . . . will do more harm than good. The proposal would indubitably mean the blocking (using this logged-in only registration) of most AOL IPs, Netscape IPs, school districts, public-use computers, and major corporations. By only allowing logged-in users on these IPs (since it is inevitable that all of them would either be blocked indefinitely or blocked consistently), in my opinion, is against the spirit of the Wiki - we're here to allow **anyone** to edit, not just those who want to create accounts. . . . This blocking policy proposal would take us one step closer to not allowing any anonymous editing - AOL, school districts, and public-use computers comprise a large amount of our editing, and many are valuable editors and contributors that we may lose if this policy is implemented. (Celloguy, 2005)

Others countered with a pragmatic argument. On the face of it, it might appear there are more restrictions as there is a new feature in the software, yet it would further the goal of greater access in practice:

I really can't figure out what you're arguing here, though. Because

right now, when an AOL IP is blocked, you can't edit using it regardless of whether or not you register. As I understand it, the proposal is to allow logged in users to edit when they otherwise wouldn't. Sure, this might lead to admins being more liberal with IP blocks, but it doesn't require it - whether or not admins are more liberal with IP blocks is a separate issue, and we could pass policies to ensure that this doesn't happen. (DiPierro, 2005a)

Much to the chagrin of the community, this proposal gained major attention with the publication of a *New York Times* article entitled “Growing Wikipedia Revises Its ‘Anyone Can Edit’ Policy” (Hafner, 2006). This led Wales to comment that:

... not every case of allowing more people to edit would count as “more open”. For example, if we had a rule that “Only Jimbo is allowed to edit this article” then this would be a lot LESS open than “no one is allowed to edit this article”. Openness refers not only to the number of people who can edit, but a holistic assessment of the entire process. I like processes that cut out mindless troll vandalism while allowing people of diverse opinions to still edit. Those are much better than full locking. (Wales, 2006c)

On June 21 the *New York Times* corrected its original article by noting that some form of blocking had always existed on Wikipedia, and the online version's headline now reads “Wikipedia refines” its policy, rather than “revises” it (Hafner, 2006). In July of 2006 acceptance of the blocking proposal was characterized as an “avalanche” of support (Wikipedia, 2006p) and I have seen little evidence that it has negatively affected users so far.

How does this story of anonymous users, vandals, and blocking engage my earlier notion of an open content community? I think there are four issues worth explicitly identifying so as to answer this question.

First, what is the scope of “anyone”? Does “anyone” include persistent vandals with no goal other than disturbing Wikipedia? The community has comfortably concluded that it does not—though it does continue to be quite

forgiving by preferring suspension and a process of escalation before outright banning occurs. Does “anyone” include anonymous editors? Historically it has, and continues to do so except in cases of suspected abuse.

Second, how to balance values? Openness is not the only value of Wikipedia, it is not even the primary one. The ultimate goal of Wikipedia is to produce a high quality encyclopedia. Many believe openness furthers the ultimate goal of producing quality content, but a quality encyclopedia should not be sacrificed in the face of a detrimental openness. Fortunately, the values of openness, quality, and kindness are often seemingly sympathetic to each other. Yet, as seen, there are cases in which they are in tension (Nagel, 1979) and require balancing, sometimes through technical intervention (Flanagan et al., 2006, p. 9). Sanger, with the Citizendium project, for example, has chosen a different balance by requiring all contributors to use real-world identities.

Third, does *possibly* imply *essentially*? In a variation of an argument by Langdon Winner (1986) wherein certain technologies (e.g. nuclear) can be inherently political (i.e., inherent to certain social and political relationships), some critics maintain that because certain things are *possible* on Wikipedia they are *essential* to Wikipedia. Whereas Winner argues that the dominant uses of the technology shape its relation to the social, Wikipedia critics argue that even a possibility is determining, or to put it another way, “because Wikipedia permits foo, it is foo’ish.” Others respond that marginal cases do not define the whole and should not be catered to if they conflict with more central values. To this end, Wales was quoted in the *New York Times* article as saying: “Protection is a tool for quality control, but it hardly defines Wikipedia. What does define Wikipedia is the volunteer community and the open participation” (as cited in Hafner, 2006).

Fourth, do technological constraints always imply movement away from openness? The ability to block anonymous users associated with an abusive IP

number was a new feature. Yet, innocent anonymous users would have been blocked before, as would have those users signed in at that IP address. With the new feature the latter group has access it did not before. In this case we see the relevance of historical context (existing practice) and practical effect on the meaning of a technical feature (Woolgar, 1991; Friedman and Nissenbaum, 1996). Consequently, “We must also study the complex interplay between the system or device, those who built it, what they had in mind, its conditions of use, and that natural, cultural, social, and political context in which it is embedded. . .” (Nissenbaum, 2001, p. 120).

Ultimately, how does this case bear upon my specification of an open content community? I believe the relevant criterion is that of nondiscrimination which, “prohibits arbitrary discrimination against persons, groups, or characteristics not relevant to the community’s scope of activity.” Although some might argue any effort to block even problematic users is a step away from openness, a chaotic culture of undisciplined vandals would equally disenfranchise those who wish to make a positive contribution.

Open Communities and Closed Law

In June of 2006 I noted, what I thought to be, three important stories for open content communities. The *Wikipedia Signpost*, Wikipedia’s local “newspaper,” reported on a new “oversight” feature that permits hidden edits to be made to Wikipedia. In the same month, it also reported that the Wikimedia Foundation had hired the same person as “general counsel and interim executive director” (Wikipedia, 2007bp). It was also at this time that it appeared the Debian Project, a GNU/Linux distribution, might sever its relationship with its legally chartered non-profit organization (Goerzen, 2006). What do these events have in

common? They demonstrate that open content communities with a formal legal standing are a conflicted beast.

Wikipedia's success has led to a difficult challenge of interfacing with the external world. As Wikipedia has grown in size and repute the likelihood of Wikipedia being subject to legal action has similarly grown. The tension between openness and closeness in such a community is no better demonstrated than by the WP:Office action (Wikipedia, 2006aa), the removal of illegal or risky material without public discussion or revision.

On Wikipedia one is expected to discuss the editing of an article with fellow contributors. Arguments are made in the open with reference to public policies. However, for those with a proprietary interest, this process of reasoned discussion can be circumvented via a call or letter to the "Wikipedia office," that is, formally contacting the Wikimedia Foundation. And, sometimes, rightfully so. What obligation did Seigenthaler (2005), someone completely unfamiliar with wikis, have to edit Wikipedia in order to remove the libelous claim that he was implicated in the assassination of the Kennedys? None. As Wales (2006d) wrote, "The problem we are seeing, again and again, is this attitude that some poor victim of a biased rant in Wikipedia ought to not get pissed and take us up on our offer of 'anyone can edit' but should rather immerse themselves in our arcane internal culture until they understand the right way to get things done."

However, unfortunately, the office mechanism can be abused by those pushing a non-encyclopedic point of view (POV), such as promoting (or censoring negative views of) a commercial product. If such people can't win their arguments on the merits of notability and neutrality within the community, having their lawyer call the office might prompt an office intervention—such as blanking or deleting the contentious article which would then be labeled with the WP:Office tag.

Something like WP:Office was an unfortunate though (probably) necessary mechanism whereby reasoned discussion is excepted so as to avoid legal problems. Yet, in an ironic twist, WP:Office soon became a red flag to those who dislike this intervention or otherwise like to make trouble for Wikipedia (e.g., copying sensitive or contentious materials off Wikipedia to continue a controversy). Whereas office actions were intended to quickly and quietly remove a potential liability, they became a flash-point. This led to the sad case in which office actions were taken without being labeled as such and a “good-faith” administrator was desysopped and blocked indefinitely because he had reverted the hidden landmine of an unlabeled office action. (Fortunately, his response (Moeller, 2006a) was an exemplar of Wikipedia tact and his position was soon restored.)

The realities of this tension between open collaboration and legal action were indicated by two Wikipedia announcements in 2006: the appointment to a CEO position (Foundation, 2006) of someone who will also act as general counsel, and the deployment of an “oversight” (revision hiding) feature (Wikipedia, 2007bj) which permits edits to be hidden from an article’s history page. Legal threats are clearly a top priority for Wikipedia.

The organizational scholar Edgar Schein (2004) argues that organizations are shaped by the crises they face in interaction with the external environment and how those events are internally integrated within the organization. This integration is not always smooth or successful, particularly for an open community. Another example of this has been a dispute in a different open content community, the Debian GNU/Linux distribution project. In a thread entitled “Who Can Make Binding Legal Agreements” (Goerzen, 2006), the Debian community argued over the licensing implications of Sun’s Java. For the sake of convenience, sometimes free software distributions will provide (controversial) “non-free”

repositories of popular applications that are often free to use (e.g., the Java development language, or the Opera browser) but do not satisfy the Free Software Definition. Sun had been tinkering with making Java more acceptable to the FOSS community, and with the help of some of Debian's leadership, Java was to be included in the non-free repositories. Other Debian contributors and Software in the Public Interest (SPI), Debian's legal parent organization, objected because of possible indemnity problems. Part of the discussion resolved around the questions of authority: community and legal. Can SPI preempt Debian leadership or decision making processes? Can Debian decisions foist liability upon SPI that SPI is unwilling to accept? In the current legal environment these are difficult issues for the type of open collaboration that Wikipedia and Debian represent.² Such concerns over copyright and liability have led Larry Sanger (2005b) to coin the neologism of "shopworks," for works that are "developed in a strongly collaborative way," and to argue they merit special protection under a law that is sensitive to the novel way in which they are produced.

In any case, what are the implications of organizational maturation on openness? The sociologist Max Weber made an important observation of how leadership often shifts from a charismatic leader to a more bureaucratic form of governance as a community matures (Weber 1978, p. 212-302; Mommsen 1992, p. 42). Clay Shirky, a contemporary scholar of organizations, has observed that "Process is an embedded reaction to prior stupidity," meaning "an organization slowly forms around avoiding the dumbest behaviors of its mediocre employees, resulting in layers of gunk that keep its best employees from doing interesting work..." (Shirky, 2003b).

Wikis do not add unnecessary process in and of themselves: they are simple, accessible, flexible, quick, and cumulative. Furthermore, community

²Though it took some time, Sun eventually released Java under the free software GPL license, making its inclusion in Debian much simpler.

process need not be overly specified in fear of a mistake since content changes are easily reverted. (The role of leadership relative to over-specifying process is discussed further in chapter 6.) However, an unforeseen implication of the wiki's ability to facilitate content creation is that policies are but another type of content. So, in the end, Wikipedia is no exception. In fact, despite the Wikipedia norm of "Avoid Instruction Creep" (Wikipedia, 2007ba) it seems unavoidable. For example, Andrew Lih, Wikipedia administrator and journalist, prompted a discussion over what he saw as an overly officious statement warning of the "speedy deletion" of a page he found useful. The deletion notice warned that the article, "is a very short article providing little or no context (CSD A1), contains no content whatsoever (CSD A3), consists only of links elsewhere (CSD A3) or a rephrasing of the title (CSD A3)" (Lih, 2007). Lih responded:

It's incredible to me that the community in Wikipedia has come to this, that articles so obviously 'keep' just a year ago, are being challenged and locked out. . . . It's as if there is a Soup Nazi culture now in Wikipedia. There are throngs of deletion happy users, like grumpy old gatekeepers, tossing out customers and articles if they don't comply to some new prickly hard-nosed standard. It's like I'm in some netherworld from the movie *Brazil*, being asked for my Form 27B(stroke)6. (Lih, 2007)

Some degree of policy is necessary in any community, and bureaucratization is a common—many would say unavoidable—feature of organizational development. And even in the face of a proliferation of process the open content community criteria of transparency and integrity is largely preserved. However, should the accretion of policy become too heavy integrity can become compromised by frustration and "Wikilawyerling" (employing overly technical or legalistic arguments that focus on the letter of policy rather than its spirit) (Wikipedia, 2007bo). For example, the policy boom has prompted one Wikipedian to declare that he had "Kicked the Process Habit": "So as of today, I'm just going

to go ahead and edit. Lord knows the rules are making me nervous and depressed. So I'll follow all the stuff I can remember, and not try too hard to learn the other stuff" (Sandifer, 2006).

Finally, legal requirements for and upon an accountable agent of an open content community are problematic, no doubt. Actions taken outside the review of the community fall short on the criteria of transparency, and perhaps integrity. Hopefully, open content communities might bear this incongruity if incidents prompting such action remain rare and marginal.

Enclaves and Gender

One should not be surprised that a source of contention in open content communities is when a subset of community members create a closed space. The conditions that prompt such proposals and the rhetoric marshaled to support or attack them give insight into a community's attempts to understand and implement openness.

A common feature of online communities operating under an ethos of open and egalitarian values is frustration with the co-existence of group decision-making and seemingly contrary forms of autocratic authority. Evidence of this phenomenon includes the alleged "secret cabal" of USENET in the 1980s (Pfaffenberger, 1996), private "sysop" only e-mail lists or IRC channels, and "benevolent dictators" (Wikipedia, 2006b) of communities including Python, Linux, and Wikipedia. For example, consider the following comment on a Wikipedia e-mail list:

There are many private, semi-private and secret lists in which wikimedians make decisions with each other without ever telling anyone or explaining. Openness has gone overboard a very long time ago. Most things you read on the public lists have been discussed privately long before an outsider found out about them.

(van Kalken, 2006)

In a subsequent chapter, I argue that the open character of such communities fosters the emergence of strong leaders who make autocratic decisions when the community can not easily make the decision itself. Furthermore, such leadership must be exercised with humility and humor or the community may fork or fail. However, in this chapter I concern myself with another difficult situation for an open content community: the creation of an enclave (Sunstein, 2003, p. 158), or minority-specific space. For while cabal formation is a seemingly inevitable structural result of group decision making, and “closed law” is an inescapable reality of living in a litigious society, enclaves are purposely chosen by a subset of the community in seeming contradiction with the values of openness and equality. This was aptly demonstrated in the Wikipedia community by the announcement (Angela, 2006) of a “WikiChix” list for female only discussion:

Offlist chat about the recent discussions on systemic gender bias in Wikipedia made it clear that a number of women were not comfortable contributing to the conversation there. This inspired the creation of WikiChix in November 2006. WikiChix is a wiki and mailing list for female wiki editors to discuss issues of gender bias in wikis, to promote wikis to potential female editors, and for general discussion of wikis in a friendly female-only environment. (WikiChix, 2007)

Formally excluding anyone from the larger community prompts questions of fairness and discrimination. Some members reacted by arguing of a slippery slope towards absurdity, such as a need for “a mailing list for homosexual African-Americans from planets other than earth” (Alphax, 2006). In a similar spirit, another Wikipedian asks about the need for a “British-only or atheist-only” list but also acknowledges the specific motivations for the creation of Wikichix: “the list was organised to avoid a specific problem - women feeling uncomfortable posting to this male-dominated list where explicitly sexist statements (even if they weren’t

meant seriously) are left unchallenged by a large number of people” (Guettarda, 2006a).

In fact, the very notion of equality may inhibit constructive action towards mitigating bias. After interviewing male and female students about computer usage and its larger culture (i.e. reading computer magazines) Fiona Wilson (2003) argues that women who might otherwise object to informal bias might simply accept the presumption of equality or not want to challenge it so as to avoid being singled out. The model of female “chix” projects (e.g., LinuxChix, Ubuntu Women, Debian Women, KDE Women, WikiChix, etc.) appears to be a positive counterforce to this tendency.

Another response employed by those concerned with such spaces is not to object to the exclusion, but to the division of the larger community. Shouldn’t the community ensure the common space is accessible rather than spinning off groups? For example, “A better solution would be to kick any of the men that behave like that, not to assume that ‘all men are chauvinist pigs’” (Giusca, 2006). Of course, given the value accorded to free speech, the community would have a difficult time restricting the speech of “men who behave like that.” How would such a determination be made? One of the few standards available for the discrimination of speech in online communities is that of “trolling”: contentious speech, probably not even genuinely held, expressed for the sole purpose of inflaming discussion. But how would one distinguish between misogyny and trolling (Guettarda, 2006b)? (Or, how does one distinguish between genuine racism and provocation? Consider Michael Richards’, *Seinfeld*’s Kramer, 2006 claim that he is not a racist despite his racial tirade against hecklers: one need not actually believe the offensive statements one uses to antagonize others in a heated moment.) An irony is that falsely held misogynistic statements espoused for the purposes of trolling might be censured or censored, but a genuine misogynist

could claim that any formal censure is a form of “thought crime” which is generally anathema under free speech principles.

This type of discussion that traverses the difficult questions of freedom and equality often prompt extensive debate. Although discussion about these values sometimes creates a shared “productive ethical orientation” within the community (Coleman, 2005), they can also be alienating and seemingly endless. This is why such topics are so suitable to trolling in the first place, and for which community leaders often step in:

The point is, if the broad philosophical question is “Do we ban people for merely holding unpleasant or unpopular beliefs?” then the answer is “no, we never have, and there seems to be very little support for doing so”. If the point is “Does asserting unpleasant or unpopular beliefs automatically get you a free pass to be any sort of jerk you like, because we are planning to bend over backwards to make sure we don’t ever ever ever discriminate against Nazis?” then the answer is, “no, being a disruptive troll is still being a disruptive troll.” (Wales, 2006b)

Not surprisingly, it did not take long for the WikiChix proposal to be challenged; a longtime male contributor and self-described “overly combative” “anarchist” (Wikipedia, 2007ar) tried to subscribe to the list and was rejected. (I suppose this action was a violation of the norm “Do Not Disrupt Wikipedia to Illustrate a Point” (Wikipedia, 2006t), which brings some measure of sanity to difficult issues.)

The final, parliamentary, objection to the WikiChix proposal was that this exclusive list was being hosted by the Wikimedia Foundation. The other free software related women fora, while focused on being “women-friendly,” are more or less open and affiliated with the larger community. LinuxChix “is intended to be an inclusive group where everyone is and feels welcome. . . . LinuxChix is intended to be primarily for women. The name is an accurate reflection of that fact. Men are welcome because we do not want this group to be

exclusive” (Vesperman and Richardson, 2002). Debian Women states: “We’re not segregated. Debian Women is a subgroup of Debian that allows anyone to join and help” (Clark, 2005). On UbuntuWomen, “Membership is open to all” (Project, 2006). The KDE Women website is run by women so “you have to be a woman” (Webmaster, 2007), but in addition to the six listed female members, there are also five male “supporters” and men are present on the IRC channel and mailing list. The gender exclusivity of WikiChix is atypical and it is not clear to what extent this decision was considered, purposeful, and what the consequences might be relative to the other female friendly fora.

In the end, the WikiChix list was moved from being hosted by Wikimedia, which might carry the presumption of endorsing exclusive discrimination, to a non-Wikimedia host: “Excellent. I still think it’s a bad idea, but if it’s not being supported in any way by Wikimedia Foundation there’s no need to complain about it here any more” (Derksen, 2006). As is often the case on difficult issues, the conclusion to this argument was facilitated as much by exhaustion as by reason. Endless argument about whether bias exists, rather than partaking in constructive dialogue on how to counter it, is a reason such spaces are often created. By severing any support and official affiliation with the Wikimedia Foundation the topic became moot to the larger community.

While this particular case was resolved with a simple techno-institutional move, it still is illustrative of a challenge to openness. As Freeman (1996) noted informal—though no less exclusionary—boundaries may persist despite the absence of formal exclusions. Therefore “formal” enclaves can be a productive response to the “tyranny” of informal structures and biases of a larger community. To this end, Nancy Fraser (1992) proposes the notion of “subaltern counterpublics in order to signal that they are parallel discursive arenas where members of subordinated social groups invent and circulate counterdiscourses to formulate

oppositional interpretations of their identities, interests, and needs” (p. 123). Cass Sunstein (2003) recommends that in circumstances in which high status members dominate lower status members, “it can be indispensable to allow spaces in which members of minority groups, or politically weak groups, can discuss issues on their own. Such spaces are crucial to democracy itself” (p. 158).

Yet, WikiChix’s exclusion of males, rather than being open and pro-female, is obviously problematic with respect to the open content community criteria of transparency and nondiscrimination. Also, Sunstein (2003) recognizes enclaves can further group polarization and marginalization and recommends that enclave members be brought back into it contact with the larger community, otherwise, self insulation can yield extremism (p. 160-161). In following this issue I haven’t perceived a decrease in female presence after the provisioning of a female friendly space. A counter to the hypothesis that “women are abandoning the common space” is that “having a more supportive space to fall back upon will encourage comfort in speaking in common spaces.” Yet, these other female specific spaces are also open, whereas WikiChix is gender exclusive. In the end, time will tell, and I expect that because all constituencies still possess a common object (Wikipedia) marginalization and extremism will be minimal.

Conclusion

In this chapter I test the Wikipedia community against a normative specification of an *open content community*. While this notion was first developed in the context of FOSS, so as to be able to speak about the openness of a community rather than their copyright license, I believe it remains useful for communities that develop content beyond code, such as Wikipedia. Despite some difficulties, I believe Wikipedia largely satisfies the criteria.

I also explore three cases that challenge the “soft middle” of the criteria (i.e. transparency, integrity, and nondiscrimination). Boundaries are a fundamental feature of any community, but especially so for those that aspire to openness because it is rarely a simple binary: open or closed. As Clay Shirky (2006) writes, “successful open systems create the very conditions that require and threaten openness. Systems that handle this pressure effectively continue (Slashdot comments.) Systems that can’t or don’t find ways to balance openness and closedness—to become semi-protected—fail (Usenet.)” The question, then, is one of in what ways is a community open or to what degree? Even a theoretically perfect openness can lead to behavior and informal structures that are less than inclusive. So as to not be left with a gummy mess on the questions of transparency, integrity, and nondiscrimination, the open content community criteria do have a hard shell: products must be under an open license which also implies non-interference. Should some find themselves mired in controversies about cliques and discrimination they can copy the result of their collaborative efforts and work on it in a different setting (i.e., fork), which may be more or less open and, in time, forked again. This yields an important descriptive feature of an open content community: a lot of discussion about what it means to be open and threats of, or even instances of, forks. By this measure, Wikipedia certainly qualifies and, furthermore, is evidence of the continuance of the universal vision of increased access to knowledge and its production.

BIBLIOGRAPHY

- Aigrain, P. (2003). The individual and the collective in open information communities. *16th BLED Electronic Commerce Conference*, 9-11.
- Alphax (2006). Re: Fwd: Request to mailing list Wikichix-l rejected. wikien-l.
- Anderson, C. (2006). *The Long tail: why the future of business is selling less for more*. Hyperion, New York.
- AndyL (2005a). Neo-nazis attempt to stack wikivote on Jewish. wikien-l.
- AndyL (2005b). [wikien-l] A neo-nazi wikipedia. wikien-l.
- Angela (2006). Introducing a new mailing list. wikien-l.
- Anonymous et al. (2007). *How to Writer's Guide - wikiHow*. wikiHow.
- Anthere (2005a). Re: NYTimes.com: Google may host encyclopedia project. wikipedia-l.
- Anthere (2005b). Re: Sanger's memoirs. wikipedia-l.
- Anthony, D., Smith, S. W., and Williamson, T. (2005). *Explaining quality in Internet collective goods : zealots and Good Samaritans in the case of Wikipedia*. unpublished.
- Arthur, C. (2005). Log on and join in, but beware the web cults. *The Guardian*.
- Baker, N. (2008). The charms of Wikipedia. *The New York Review of Books*, 55(4).
- Ball, P. (2007). The more, the wikier: the secret to the quality of Wikipedia entries is lots of edits by lots of people. *news@nature.com*.
- Barasch, M. I. (2005). *Field notes on the compassionate life: a search for the soul of kindness*. Rodale, New York.
- Barger, J. (1997). Beyond the Interpedia? alt.internet.search.
- Barnett, C. (2005). Wiki Mania. *Florida Trend*.
- Barthes, R. (2005). *The neutral*. Columbia University Press, New York.

- Basbanes, N. A. (1999). *A gentle madness: bibliophiles, bibliomanes, and the eternal passion for books*. Henry Holt and Company, New York.
- Bass, B. M. (1990). *Bass & Stogdill's handbook of leadership : theory, research, and managerial applications*. Free Press, New York, 3 edition.
- Batson, C. D. (1991). *The altruism question: toward a social-psychological answer*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- Battles, M. (2007). Authority of a new kind. *Britannica Blog: Web 2.0 Forum*.
- Beck, K., Mike Beedle, A. v., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J., and Thomas, D. (2001). *Manifesto for Agile Software Development*.
- Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A., and Tipton, S. M. (1996). *Habits of the heart: individualism and commitment in American life*. University Of California Press, Berkeley.
- Benkler, Y. (2002). Coase's penguin, or, Linux and the nature of the firm. *The Yale Law Journal*, 112(3):369–446.
- Benkler, Y. (2006a). On "digital Maoism". *Edge*.
- Benkler, Y. (2006b). *The wealth of networks: how social production transforms markets and freedom*. Yale University Press, New Haven.
- Benkler, Y. and Nissenbaum, H. (2006). Commons-based peer-production and virtue. *The Journal of Political Philosophy*, (4).
- Berners-Lee, T. (1999). *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*. HarperSanFrancisco, San Francisco.
- Blackburn, S. (1996). Culture. In Blackburn, S., editor, *The Oxford Dictionary of Philosophy*. Oxford University Press, Oxford.
- Bohm, D. (1996). *On dialog*. Routledge, New York.
- Bosco, G. (2004). Implicit theories of "good leadership" in the open-source community. Master's thesis, Technical University of Denmark.
- Bowles, S. and Gintis, H. (1998). The moral economy of communities: structured populations and the evolution of pro-social norms. *Evolution & Human Behavior*, 19(1):3–25.
- Bowrey, K. and Matthew (2005). Rip, Mix, Burn. *First Monday*, 7(8).

- Boyd, D. (2005a). Academia and Wikipedia. *Many2Many*.
- Boyd, D. (2005b). Wikipedia, academia and Seigenthaler. *Many2Many*.
- Boyd, D. (2007). Knowledge access as a public good. *Britannica Blog: Web 2.0 Forum*.
- Briggs, J., Nunamaker, R., Mittleman, D., Vogel, D., and Balthazard, P. (1997). Lessons from a dozen years of group support systems research: a discussion of lab and field findings. *Journal of MIS*, 13:163–207.
- Britannica (1911a). Ku Klux Klan (1911). In *Encyclopedia Britannica*, volume 17. 170.
- Britannica (1911b). Lynch Law (1911). In *Encyclopedia Britannica*, volume 15. 943.
- Brody, H. (1997). Great expectations: why technology predictions go awry. In Teich, A. H., editor, *Technology and the Future*, chapter 9. St. Martin's Press, New York, 7 edition.
- Bryant, S. L., Forte, A., and Bruckman, A. (2005). Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*. ACM, New York.
- Bullard, F. (1968). *Volcanoes*. University of Texas Press, Austin.
- Burawoy, M. (1991). Reconstructing social theories. In *Ethnography Unbound: Power and Resistance in Modern Metropolis*, chapter 2, pages 8–27. University Of California Press, Berkeley.
- Buriol, L. S., Castillo, C., Donato, D., Leonardi, S., and Millozzi, S. (2006). Temporal analysis of the wikigraph. In *Web Intelligence Conference 2006*.
- Burke, P. (2000). *A social history of knowledge: from Gutenberg to Diderot*. Polity Press, Cambridge.
- Burrell, G. and Morgan, G. (1979). *Sociological paradigms and organizational analysis*. Heinemann, London.
- Bush, V. (1945). As we may think. *Atlantic Monthly*.
- Capocci, A., Servedio, V. D. P., Colaioni, F., Buriol, L. S., Donato, D., Leonardi, S., and Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The Internet encyclopedia Wikipedia. *The American Physical Society*, 74(3).

- Carr, N. (2005). The amorality of Web 2.0. *Rough Type*.
- Carr, N. (2007). Stabbing Polonius. *Rough Type*.
- Celloguy, F. (2005). Re: Blocking proposal. wikien-l.
- Ceruzzi, P. (1997). An unforeseen revolution: computers and expectations, 1935-1985. In Teich, A. H., editor, *Technology and the Future*, chapter 10. St. Martin's Press, New York, 7 edition.
- Chopra, S. and Dexter, S. (2007). *Decoding liberation: the promise of free and open source software*. Routledge, New York.
- Ciffolilli, A. (2003). Phantom authority, self-selective recruitment and retention of members in virtual communities: the case of Wikipedia. *First-Monday*, 8(12).
- Clark, E. (2005). Frequently Asked Questions. Debian Women.
- Coffin, J. (2006). Analysis of Open Source principles and diverse collaborative communities. *First Monday*, 11(6).
- Coleman, G. (2005). Three ethical moments in Debian: the making of an (ethical) hacker, part III. In *the social construction of freedom in free and open source software: actors, ethics, and the liberal tradition*.
- Conservapedia (2007). Examples of Bias in Wikipedia. Conservapedia.
- Coulon, A. (1995). *Ethnomethodology*, volume 36 of *Qualitative Research Methods*. Sage, Thousand Oaks.
- Cramton, C. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science*, 12:346–371.
- Cunningham, W. (2005). Wikis then and now. In *Keynote at Wikimania 2005*.
- Darnton, R. (1979). *The business of enlightenment: a publishing history of the Encyclope'die*. The Belknap Press of Harvard University, Cambridge.
- David, S. (2004). Opening the source of accountability. *First Monday*, 9(11).
- Deetz, S. (1996). Describing differences in approaches to organization science: rethinking Burrell and Morgan and their legacy. *Organization Science*, 7(2).
- Denning, P., Horning, J., Parnas, D., and Weinstein, L. (2005). Inside risks: Wikipedia risks. *Communications of the ACM*, 48(12):152.

- Derksen, B. (2005). Re: [wikien-l] Re: Article deletion (proposed 1 month hiatus). wikien-l.
- Derksen, B. (2006). Re: Fwd: Request to mailing list Wikichix-l rejected. wikien-l.
- Devouard, F. (2007). Welcome Mike ! *Foundation-l*.
- Dewey, M. (2004). *A classification and subject index for cataloguing and arranging the books and pamphlets of a library (1876)*. Project Gutenberg.
- dharma (2004). Agreed. *Kuro5hin*.
- Dibona, C., Ockman, S., and Stone, M. (1999). *Open Sources: voices from the Open Source revolution*. O'Reilly, Cambridge.
- Dickens, D. (2004). *The Tale of Two Cities: A community reading project: notes on issue 1: glossary part 2 of 5*. Stanford, Standford.
- DiPierro, A. (2005a). Re: Blocking proposal. wikien-l.
- DiPierro, A. (2005b). Re: Quality vs. consensus (was: Meanwhile, AfD grinds on). wikien-l.
- Doctorow, C. (2006). On "digital Maoism". *Edge*.
- Drum, K. (2004). Google and the human spirit: a reply to Michael Gorman. *Washington Monthly*.
- Duguid, P. (2006). Limits of self-organization: peer production and laws of quality. *FirstMonday*, 11(10).
- Dyer, S. (1984). Re on the direction of net.motss. *net.motss*.
- Einbinder, H. (1964). *The myth of the Britannica*. Grove Press, New York.
- Eisenstein, E. L. (1993). *The printing revolution in early modern Europe*. Cambridge University Press, Cambridge, canto edition.
- Elliott, M. (2006). Stigmergic collaboration: the evolution of group work. *M/C Journal*, 9(2).
- Encarta (2006). Encyclopedia. In *Microsoft Encarta Online Encyclopedia*. Microsoft Corporation.
- Encyclopedia, D. (1999). *The Distributed Encyclopedia introduction*.

- Engelbart, D. C. (1962). *Augmenting human intellect: a conceptual framework*. Menlo Park, Ca.
- Etzioni, A. (1975). *Comparative analysis of complex organizations*. Free Press of Glencoe, NY, revised edition.
- Flanagan, M., Howe, D., and Nissenbaum, H. (2006). Embodying values in technology: theory and practice. In Jeroen van den Hoven, J. W., editor, *Information Technology and Moral Philosophy*. Cambridge University Press, Cambridge.
- Fogel, K. (2005). *Producing open source software: how to run a successful free software project*.
- Fontaine, H. L. and Otlet, P. (1990). Creation of a universal bibliographic repertory: a preliminary note (1895-1896). In Rayward, W. B., editor, *International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Elsevier, Amsterdam.
- Forrester, J. D. (2005). Re: Your golorous [sic] leader. wikien-l.
- Forte, A. and Bruckman, A. (2005). Why do people write for Wikipedia? Incentives to contribute to open-content publishing. In *GROUP workshop: Sustaining community: The role and design of incentive mechanisms in online systems*.
- Foucault, M. (1984). What is an author? In Rabinow, P., editor, *The Foucault Reader*. Pantheon Books, New York.
- Foundation, W. (2005). *Board of Trustees*.
- Foundation, W. (2006). *Resolution CEO*.
- Foundation, W. (2007a). *Planet Wikimedia*.
- Foundation, W. (2007b). Wikipedia Reaches 2 Million Articles. Wikimedia Foundation.
- Fraser, N. (1992). Rethinking the public sphere: a contribution to the critique of actually existing democracy. In *Habermas and the public sphere*, pages 109–142. The MIT Press, Cambridge.
- Freeman, J. (1996). *The tyranny of structurelessness: why organisations need some structure to ensure they are democratic (1970)*. Struggle.
- Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions in Information Systems*, 14(2):330–346.

- froosh (2006). Main Street Meets Madison Avenue, Wall Street and Silicon Valley – what is Wikipedia.org’s valuation? *HipMojo.com*.
- Galinsky, A. and Mussweiler, T. (2001). First offers as anchors: The role of perspective taking and negotiator focus. *Journal of Personality and Social Psychology*, 81.
- Garfield, E. (1953). *Librarian versus documentalist*. Unpublished.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Prentice-Hall, New Jersey.
- Garzarelli, G. and Galoppini, R. (2003). Capability coordination in modular organization: Voluntary FS/OSS production and the case of Debian GNU/Linux. In *Economics Working Paper Archive at WUST*, number 0312005. Industrial Organization.
- Geertz, C. (1973). Thick description: toward an interpreted theory of culture. In *The Interpretation of Cultures and Local Knowledge*, chapter 1, pages 3–30. Basic Books, New York, NY, clifford geertz edition.
- GeorgeLouis (2006). Wikipedia talk:neutral point of view. Wikimedia.
- Giusca, B. (2006). Re: Fwd: Request to mailing list Wikichix-l rejected. wikien-l.
- Glaser, B. and Strauss, A. (1967). *The discovery of grounded theory: strategies for qualitative research*. Aldine Publishing Company, Chicago.
- Godwin, M. (1994a). Meme, Counter-meme. *Wired*, 2(10).
- Godwin, M. (1994b). Nine principles for making virtual communities work. *Wired*, 2(6).
- Goerzen, J. (2006). *Who can make binding legal agreements*.
- Golden-Biddle, K. and Locke, K. (1993). Appealing work: An investigation of how ethnographic texts convince. *Organization Science*, 4:595–616.
- Goldschmidt, R. B. and Otlet, P. (1990). On a new form of the book: the microphotographic book (1906). In Rayward, W. B., editor, *International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Elsevier, Amsterdam.
- Gorman, M. (2004). Google and God’s mind: the problem is, information isn’t knowledge. *Los Angeles Times*.
- Gorman, M. (2005). Revenge of the blog people! *Library Journal*.

- Gorman, M. (2007a). Jabberwiki: the educational response, part I. *Britannica Blog: Web 2.0 Forum*.
- Gorman, M. (2007b). Jabberwiki: the educational response, part II. *Britannica Blog: Web 2.0 Forum*.
- Gorman, M. (2007c). The siren song of the Internet: part I. *Britannica Blog*.
- Gorman, M. (2007d). The siren song of the Internet: part II. *Britannica Blog*.
- Gorman, M. (2007e). *Web 2.0: the sleep of reason, part II*.
- Gould, S. J. (2000). Deconstructing the "science wars" by reconstructing an old mold. *Science*, 287(5451):253–261.
- Gouthro, L. (2000). Building the world's biggest encyclopedia. *CNN*.
- Gove, P. (1961). *Webster's Third New International Dictionary, Unabridged*. Merriam-Webster.
- Graham, P. (2005). *Web 2.0*.
- Gray, A. (2005). Re: Librarians, Professors, and Pundits. Wikimedia-l.
- Guardian, T. (2005). In praise of ... the Wikipedia. *The Guardian*.
- Guettarda (2006a). Re: Fwd: Request to mailing list Wikichix-l rejected. wikien-l.
- Guettarda (2006b). Re: Fwd: Request to mailing list Wikichix-l rejected. wikien-l.
- Hafner, K. (2006). Growing Wikipedia Refines Its "anyone Can Edit" Policy. *New York Times*.
- Hart, M. (1992). History and philosophy of Project Gutenberg. Project Gutenberg.
- Hart, M. (2006). Historical draft on digital encyclopedias. *Private correspondence*.
- Hawking, S. (2002). *On the shoulders of giants: The great works of physics and astronomy*. Running Press, Philadelphia.
- Headrick, D. (2000). *When information came of age*. Oxford University Press, Oxford.
- Hedrick, C. (1992). 386bsd vs Linux. *comp.os.linux*.
- Helm, B. (2005). Wikipedia: "a Work in Progress". *BusinessWeekonline*.

- Hitchens, H. (2005). *Defining the world: the extraordinary story of Dr. Johnson's dictionary*. Farrar, Straus, and Giroux, New York.
- Hollander, E. P. (1960). Competence and conformity in the acceptance of influence. *Journal of Abnormal and Social Psychology*, 61(3):365369.
- holotone (2006). Wikipedia Showdown! *grupthink*.
- House, H. (2006). *The dictionary wars II: Webster vs. Worcester*. History House.
- Howard, R. M. (1999). *Standing in the shadow of giants: plagiarists, authors, collaborators*. Ablex Publishing Corporation, Stamford.
- Jacobs, A. J. (2004). *The-know-it-all*. Simon and Schuster, New York.
- Jacobs, A. J. (2005). Wikiworld: the experiment. *Esquire*.
- Jarvenpaa, S. and Leidner, D. (1999). Communication and trust in global virtual teams. *Organization Science*, 10(6):791.
- Jashemski, W. F. (1999). Pliny the Younger's Letters 3.5.7-17 (to Baebius Macer). In *A Pompeian herbal*.
- Jazi, P. (1994). On the author effect: contemporary copyright and collective creativity. In Martha Woodmansee, P. J., editor, *The construction of authorship: textual appropriation in law and literature*. Duke University Press, Durham.
- Jefferson, T. (1813). *To Isaac McPherson Monticello*.
- Jenkins, H. (1992). *Textual poachers: television fans and participatory culture*. Routledge, New York.
- Jenkins, H. (2006). Confronting the challenges of Participatory Culture: media education for the 21st century. *Confessions of an Aca/Fan*.
- Johns, A. (2001). The birth of scientific reading. *Nature*, 409(287).
- Johnson, B. K. (2007). *Wikipedia as collective action: personal incentives and enabling structures*.
- Kahneman, D. and Tversky, A. (1995). Conflict resolution: A cognitive perspective. In et al., K. A., editor, *Barriers to Conflict Resolution*, chapter 3. Norton, New York.
- Karsai, I. (2004). Self-organization and the origin of complexity. In Matt Young, T. E., editor, *Why Intelligent Design Fails*, chapter 7. Rutgers University press, New Brunswick.

- Keen, A. (2006). Web 2.0: The second generation of the Internet has arrived. It's worse than you think. *The Daily Standard*.
- Keen, A. (2007). *The cult of the amateur: how today's Internet is killing our culture*. Doubleday, New York.
- Kelly, K. (1995). *Out of control: the new biology of machines, social systems and the economic world*. Perseus Books Group, Cambridge, reprint edition edition.
- Kelly, K. (2006a). On "digital Maoism". *Edge*.
- Kelly, K. (2006b). Scan this book! What will happen to books? Reader, take heart! Publisher, be very, very afraid. Internet search engines will set them free. A manifesto. *New York Times Magazine*.
- Kelty, C. (2005). Geeks, social imaginaries, and recursive publics. *Cultural Anthropology*, 20(2):185–214.
- Kingwell, M. (1995). *A civil tongue: justice, dialogue and the politics of pluralism*. The Pennsylvania State University Press, Pennsylvania.
- Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. (2007a). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007)*.
- Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007b). He says, she says: conflict and coordination in Wikipedia. In *CHI 2007 Proceedings: Online Representation of Self*. ACM.
- Koepp, C. (2002). Making money: artisans and entrepreneurs and Diderot's Encyclopedie. In Daniel Brewer, J. C. H., editor, *Using the Encyclopedie*. Voltaire Foundation, Oxford.
- Kogan, H. (1958). *The great EB: the story of the Encyclopaedia Britannica*. The University Of Chicago Press, Chicago.
- Kollock, P. (1999). The economies of online cooperation: Gifts and public goods in cyberspace. In Smith, M. and Kollock, P., editors, *Communities in Cyberspace*. Routledge Press, London.
- Kramer, R. M. and Carnevale, P. J. (2001). Trust and intergroup negotiation. In Brown, R. and Gaertner, S., editors, *Blackwell Handbook of Social Psychology: Intergroup Relations*, pages 431–450. Blackwell Publishers, Oxford, UK.

- Krangel, E. (2007). Two attempts at opening up religion online. *NewAssignment.Net*.
- Kriplean, T., Beschastnikh, I., McDonald, D. W., and Golder, S. A. (2007). Community, consensus, coercion, control: CS*W or how policy mediates mass participation. In *GROUP'07: International Conference on Supporting Group Work 2007*, pages 167–176. ACM.
- Kropotkin, P. (1902). *Mutual aid: a factor of evolution*.
- Krupp, A. (2005). Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- Krupp, A. (2006). *The four webs: Web 2.0, digital identity, and the future of human interaction*.
- Kuznetsov, S. (2006). Motivations of contributors to Wikipedia. In *ACM SIGCAS Computers and Society*, volume 36. ACM.
- Lanier, J. (2006). Digital maoism: the hazards of the new online collectivism. *Edge*.
- Lebert, M. (2004). *Michael Hart: changing the world through e-books*. Project Gutenberg, english edition.
- Legbagede (2001). *A limited company of useful knowledge: Paul Otlet, the International Institute of Bibliography and the Limits of Documentalism*. everything2.
- Lehmann, E. (2006). Rewriting history under the dome. *Lowell Sun Online*.
- Lessig, L. (2004). *Free culture*. The Penguin Press, NY.
- Levesque, M. (2004). Fundamental issues with open source software development. *FirstMonday*, 9(4).
- Levina, N. and Vaast, E. (2003). *Collaborating across boundaries: boundary objects and spanners in practice*. Draft.
- Lih, A. (2007). *Unwanted: new articles in Wikipedia*.
- Lincoln, A. (1861). *First Inaugural Address*.
- Lio, E. D., Fraboni, L., and Leo, T. (2005). TWiki-based facilitation in a newly formed academic community of practice. In *WikiSym '05*. ACM.

- Lois, J. (1999). Socialization to heroism: individualism and collectivism in a voluntary search and rescue group. *Social Psychology Quarterly*, 62(2):117–135.
- Lottieri, C. (2005). *Vatican City as a free society: legal order and political theology*. Ludwig von Mises Institute Working Papers.
- Maack, M. N. (2004). The lady and the antelope: Suzanne Briet’s contribution to the french documentation movement. *Library Trends*, 52(4):719–747.
- Mann, T. (2007). Brave new (digital) world, part ii: foolishness 2.0? *Britannica Blog: Web 2.0 Forum*.
- Martin, K. (2007a). An example of transparency. *Nonbovine Ruminations*.
- Martin, K. (2007b). Re: Lock new article creation for three months. wikien-l.
- McArthur, T. (1986). *Worlds of reference: lexicography, learning, and language from the clay tablet to the computer*. Cambridge University Press, Cambridge.
- McCabe, J. (1947). *The lies and fallacies of the Encyclopaedia Britannica: how powerful and shameless clerical forces castrated a famous work of reference*. Haldeman-Julius Publications, Girard Kansas.
- McCabe, J. (1951). *The Columbia Encyclopedia’s crimes against the truth*. Haldeman-Julius Publications, Girard Kansas.
- McCarthy, M. (1999). It’s not true about Caligula’s horse; Britannica checked — dogged researchers answer some remarkable queries. *Wall Street Journal*, page A1.
- McHenry, R. (2007). Web 2.0: Hope or Hype? *Britannica Blog: Web 2.0 Forum*.
- McIlwaine, I. C. (2006). *Notes & Queries*. UDC Consortium.
- McKillop, A. B. (2002). *The spinster and the prophet*. 4 Walls 8 Windows, New York.
- Mclemee, S. (2007). Mass Culture 2.0. *Inside Higher Ed*.
- Meadows, A. J. (1990). Theory in information science. *Journal of Information Science*, 16(1):5963.
- Meatball (2006a). Assumegoodfaith. MeatballWiki.
- Meatball (2006b). Meatballbackgrounder. Meatball Wiki.

- Meatball (2007a). Fairprocess. Meatball Wiki.
- Meatball (2007b). *GodKing*.
- Meatball (2007c). Righttofork. Meatball.
- Merton, R. K. (1993). *On the shoulders of giants*. University of Chicago Press, Chicago.
- Michels, R. (2001). *Political Parties (1911)*. Batoche Books, Ontario.
- Millosh (2007). Aggregated aggregators. *millosch's blog*.
- Moeller, E. (2005). Re: [[talk:autofellatio]] - Poll. wikien-l.
- Moeller, E. (2006a). Indefinite block and desysopping by User:Danny. wikien-l.
- Moeller, E. (2006b). Re: "good authors". wikien-l.
- Moeller, E. (2007a). quality.wikimedia.org and wikiquality-l launched. Wikimedia.
- Moeller, E. (2007b). Rename admins to janitors. wikien-l.
- Momigliano, A. and Cornell, T. J. (2003). *patrum auctoritas*. In Simon Hornblower, A. S., editor, *The Oxford Classical Dictionary*. Oxford University Press.
- Mommsen, W. J. (1992). *The political and social theory of Max Weber: collected essays*. University of Chicago Press, Chicago.
- Montiel-Overall, P. (2005). Toward a theory of collaboration for teachers and librarians. *American Library Association*, 8.
- Morton, H. C. (1994). *The story of Webster's Third: Philip Gove's controversial dictionary and its critics*. Cambridge University Press, USA.
- Muddiman, D. (1998). The universal library as modern utopia: the information society of H. G. Wells. *Library History*, 14.
- Mugglestone, L. (2000). *Lexicography and the OED: pioneers in the untrodden forest*. Oxford University Press, Oxford.
- Nagel, T. (1979). The fragmentation of value. In *Mortal Questions*, pages 128–141. Cambridge University Press, Cambridge, MA.
- Nelson, R. (1992). patchlevel 6. *comp.os.linux*.

- Nisbett, R. and Ross, L. (1980). *Human inference*. Prentice Hall, Englewood Cliffs, NJ.
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2(2):2542.
- Nissenbaum, H. (2001). How computer systems embody values. *Computer*, 34(3):120–118.
- Novick, P. (1988). *The noble dream: the "objectivity question" and the American historical profession*. Cambridge University Press, Cambridge.
- NSK (2005). Re: Flags. wikipedia-l.
- Nupedia (2000). Nupedia.com editorial policy guidelines. 3.31.
- Oglesby, T. K. (1903). *Some truths of history: a vindication of the South against the Encyclopedia Britannica and other maligners*. The Byrd Printing Company, Atlanta.
- O'Mahony, S. and Ferraro, F. (2003). *Managing the boundary of an 'Open' project*.
- O'Mahony, S. and Ferraro, F. (2004). *Hacking alone? The effects of all online and offline participation or an open-source community leadership*.
- O'Mahony, S. and Ferraro, F. (2007). *Governance in production communities*.
- Onion, T. (2005). Congress Abandons WikiConstitution. *The Onion*, 41(39).
- Onion, T. (2006). Wikipedia celebrates 750 years of American independence. *The Onion*.
- Open (2007). *Open Wiki Blog Planet*.
- O'Reilly, T. (2005). *What is Web 2.0? Design patterns and business models of the next generation of software*. O'Reilly.
- O'Reilly, T. (2006). *Web 2.0 Compact Definition: Trying Again*.
- Orlowski, A. (2005a). There's no Wikipedia entry for 'moral responsibility'. *The Register*.
- Orlowski, A. (2005b). Wikipedia founder admits to serious quality problems. *The Register*.

- Ortega, F. and Gonzalez-Barahona, J. (2007). Quantitative analysis of the Wikipedia community of users. In *WikiSym '07*. ACM.
- Ortega, F. and Gonzalez-Barahona, J. (2008). On the inequality of contributions to Wikipedia. In *Proceedings of the 41st Hawaiian International Conference on System Sciences (HICSS 2008)*.
- orthogonal (2006). Snowspinner muses about stalking and murder. *The Wikipedia Review*.
- OSI (2006a). History of the OSI. OSI.
- OSI (2006b). The Open Source definition. 1.9.
- Otlet, P. (1989). *Trait de documentation. Le livre sur le livre: thorie et pratique (1934)*. Centre de Lecture publique de la Communaut franaise, Lige, reprint edition.
- Otlet, P. (1990a). The science of bibliography and documentation (1903). In Rayward, W. B., editor, *International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Elsevier, Amsterdam.
- Otlet, P. (1990b). Something about bibliography (1891-1892). In Rayward, W. B., editor, *International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Elsevier, Amsterdam.
- Otlet, P. (1990c). Transformations in the bibliographical apparatus of the sciences (1918). In Rayward, W. B., editor, *International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Elsevier, Amsterdam.
- PeopleProjectsAndPatterns (2007). Wikipedia. Cunningham & Cunningham.
- Pfaffenberger, B. (1996). "if I want it, it's okay": Usenet and the (outer) limits of free speech. *The Information Society*, pages 365–386.
- Pink, D. H. (2005). The book stops here. *Wired*, 13(3).
- Poe, M. (2006). The hive: can thousands of Wikipedians be wrong? How an attempt to build an online encyclopedia touched off historys biggest experiment in collaborative knowledge. *The Atlantic Monthly*.
- Pollard, D. (2005). Will that be coordination, cooperation, or collaboration? *How to Save the World*.
- Posner, R. A. (2007). *The little book of plagiarism*. Pantheon Books, New York.

- Post, D. G. (2001). *"The Free Use of our Faculties": Jefferson, Cyberspace, and the language of social life.*
- Preece, J. (2004). Etiquette, empathy and trust in communities of practice: stepping-stones to social capital. *Journal of Universal Computer Science*, 10(3).
- Preece, J. and Ghazati, K. (2001). Experiencing empathy online. In Rice, R. R. and Katz, J. E., editors, *The Internet and Health Communication: Experience and Expectations*, pages 237–260. Sage Publications, Thousand Oaks.
- Press, A. (2005). Author of false Wikipedia entry apologizes: Tennessee man says false entry on Seigenthaler part of joke on co-worker. *MSNBC*.
- Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. In *Proceedings of GROUP 2007*.
- Prodromou, E. (2007). Your paper and presentation at WikiSym. (1193147885.6121.35.camel@bear).
- Project, U.-W. (2006). Ubuntu Women. Ubuntu-Women Project.
- Proofreaders, D. (2004). Beginning proofreaders' Frequently Asked Questions. Project Gutenberg's Distributed Proofreaders.
- Proofreaders, D. (2005). Distributed Proofreaders Update. Distributed Proofreaders.
- Quiggin, J. (2006). Wikipedia and sausages. *Out Of the Crooked Timer*.
- Rafaeli, S., Hayat, T., and Ariel, Y. (2005). Wikipedians' sense of community, motivations, and knowledge building: a cross-cultural study. In *Proceedings of Wikimania 2005*.
- Raymond, E. (1997). *The cathedral and the bazaar*.
- Raymond, E. (1998). Homesteading the Noosphere. *First Monday*, 3(10).
- Rayward, W. B. (1975). *The universe of information. The work of Paul Otlet for documentation and international organisation*. International Federation of Documentation (FID).
- Rayward, W. B. (1991). The case of Paul Otlet, pioneer of information science, internationalist, visionary: reflections on biography. *Journal of Librarianship and Information Science*, 23:135–145.

- Rayward, W. B. (1994). Visions of Xanadu: Paul Otlet (1868-1944) and hypertext. *JASIS*, 45:235–250.
- Rayward, W. B. (1997a). The origins of information science and the International Institute of Bibliography. *Journal of the American Society for Information Science*, 48:289–300.
- Rayward, W. B. (1997b). Paul Otlet and his successors the origins of Information Science and the International Institute of Bibliography/International Federation for Information and Documentation (fid). *Journal of the American Society for Information Science*, 48:289–300.
- Rayward, W. B. (1999). H.G. Wells's idea of a World Brain: a critical reassessment. *Journal of the American Society for Information Science*, 50:575–579.
- Reagle, J. (1999). Internet quotation appendix. Working draft, Berkman Center for Internet and Society, Harvard Law School.
- Reagle, J. (2004). Open content communities. *M/C: A Journal of Media and Culture*, 7.
- Reagle, J. (2005a). A case of mutual aid: Wikipedia, politeness, and perspective taking. In *Proceedings of Wikimania 2005*.
- Reagle, J. (2005b). *Four short stories about the reference work*.
- Reagle, J. (2005c). Re: Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- Reagle, J. (2005d). Re: Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- Reagle, J. (2005e). Re: Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- Reagle, J. (2006a). Is the Wikipedia neutral? In *Proceedings of Wikimania 2006*.
- Reagle, J. (2006b). Notions of openness. In *FM10 Openness: Code, Science, and Content: Selected Papers from the First Monday Conference*, volume 11.
- Reagle, J. (2006c). Open communities and closed law. In Lipika Bansal, Paul Keller, G. L., editor, *In the Shade of the Commons -Towards a Culture of Open Networks*, pages 165–1677. Waag Society Amsterdam.
- Reagle, J. (2007a). Do as I do: authorial leadership in Wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*. ACM Press, New York, NY, USA.

- Reagle, J. (2007b). Equality, gender, and speech in open communities. *Re-public*.
- Reagle, J. (2007c). Method. In *Open Codex*.
- Reed Jr., J. W. (1962). Noah Webster's debt to Samuel Johnson. *American Speech*, 37(2):95–105.
- Reinhardt, C. (2003). Collaborative knowledge creation in virtual communities of practice. In *Masters of Social and Economic Sciences*. Department of Value-Process Management, Marketing University of Innsbruck, Innsbruck.
- Reynolds, G. (2006). *An army of Davids: how markets and technology empower ordinary people to beat big media, big government, and other goliaths*. Nelson Current, Nashville.
- Rheingold, H. (2002). *Smart mobs*. Perseus Publishing, Cambridge, Mass.
- Rick (2005). Re: Re: Writing about sexual topics responsibly is not. wikien-l.
- Rockwell, G. (1999). Diderot and hypertext; dreaming the circle of knowledge. McMaster Association for 18th-Century Studies.
- Online (2006). The Wikipedia Ombudsman. wikien-l.
- Rosenberg, S. (2007). *Dreaming in code: two dozen programmers, three years, 4732 bugs, and one quest for transcendent software*. Crown Publishers, New York.
- Roth, C. (2007). Viable wikis: struggle for life in the wikisphere. In *WikiSym '07*. ACM.
- Rushkoff, D. (2003). *Open Source Democracy: How online communication is changing offline politics*. Demos.
- Sahlins, M. (2004). *Apologies to Thucydides: understanding history as culture and vice versa*. The University Of Chicago Press, Chicago.
- Saintonge, R. (2005). Re: Wikimedia Foundation Internal Radio. wikipedia-l.
- Sandifer, P. (2006). I've Kicked the Process Habit. wikien-l.
- Sanger, L. (2000a). 602 members. nupedia-l.
- Sanger, L. (2000b). *Epistemic Circularity*.
- Sanger, L. (2000c). I am a clueless newbie. nupedia-l.

- Sanger, L. (2000d). Introduction. nupedia-l.
- Sanger, L. (2000e). Re: Nupedia: Questions. nupedia-l.
- Sanger, L. (2001a). Britannica or Nupedia? The future of free encyclopedias. *Kuro5hin*.
- Sanger, L. (2001b). Interpedia is dead—long live Wikipedia. comp.infosystems.interpedia.
- Sanger, L. (2001c). Let's make a wiki. Nupedia.
- Sanger, L. (2001d). Re: Admins (i suspect LMS) permanently deleting things. wikipedia-l.
- Sanger, L. (2001e). Wikipedia is now useful! Wikipedia.
- Sanger, L. (2005a). The early history of Nupedia and Wikipedia: a memoir. *Slashdot*.
- Sanger, L. (2005b). Why collaborative free works should be protected by the law. *Slashdot*.
- Sanger, L. (2006a). Text and collaboration, part I. *Kuro5hin*.
- Sanger, L. (2006b). Toward a new compendium of knowledge. Citizendium.
- Sanger, L. (2007a). How's the unforking going? *Citizendium Forums*.
- Sanger, L. (2007b). Our gift to the world: CC-by-sa. *Citizendium Blog*.
- Sarno, D. (2007). Wikipedia wars erupt. *Los Angeles Times*.
- Schein, E. H. (1992). *Organizational culture and leadership*. Jossey-Bass Publishers, San Francisco, 2nd edition.
- Schein, E. H. (2004). *Organizational culture and leadership*. John Wiley & Sons, San Francisco, 3rd edition.
- Schiff, S. (2006). Know it all: can Wikipedia conquer expertise? *The New Yorker*.
- Schon, D. (1983). *The reflective practitioner: how professionals think in action*. Basic Books, New York.
- Schroer, J. and Hertel, G. (2007). *Voluntary engagement in an open web-based encyclopedia: Wikipedians, and why they do it*.

- Schwab, R. N. (1963). Introduction. In Schwab, R. N. and Rex, W. E., editors, *Preliminary Discourse to the Encyclopedia of Diderot*. ITT Bobbs-Merrill, Indianapolis.
- Seigenthaler, J. (2005). A false Wikipedia 'biography'. *USA Today*.
- Service, O. W. (1995). *Objectivism-related e-mail lists*.
- Shannon, P. (2000). *Regarding Sanger and Shannon's Review of Y2K News Reports*.
- Shannon, R. (2005). Re: Re: How to join the cabal. wikien-l.
- Shapin, S. (1994). *A social history of truth: civility and the science in seventeenth-century England*. The University Of Chicago Press, Chicago.
- Sheeran, M. (1996). *Beyond majority rule: voteless decisions in the Religious Society of Friends*. Philadelphia Yearly Meeting, Philadelphia, PA.
- Shimo, A. (2006). Prolific Canadian is king of Wikipedia. *globeandmail.com*.
- Shirky, C. (2003a). A group is its own worst enemy. In *O'Reilly Emerging Technology Conference*.
- Shirky, C. (2003b). Wikis, grafitti, and process. *Many-to-Many*.
- Shirky, C. (2005). Wikipedia: me on Boyd on Sanger on Wales. *Many2Many*.
- Shirky, C. (2006). News of Wikipedia's death greatly exaggerated. *Many-to-Many*.
- Shirky, C. (2007a). *Gorman, redux: the siren song of the Internet*.
- Shirky, C. (2007b). Old revolutions good, new revolutions bad: a response to Gorman. *Many2Many*.
- Sj (2005). Re: Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- SlimVirgin (2005). Re: Original Research versus Point of View,. wikien-l.
- Smith, M. (1994). *Social science in the crucible: the American debate over objectivity and purpose, 1918 – 1941*. Duke University Press, Durham.
- Sproull, L. (2003). Online communities. In *The Internet Encyclopedia*, pages 733–744. John Wiley, New York, hossein bidgoli (ed.) edition.

- Sproull, L., Conley, C. A., and Moon, J. Y. (2004). Prosocial behavior on the Net. In Amichahi-Hamburger, Y., editor, *The Social Net: the Social Psychology of the Internet*. Oxford University Press, Oxford.
- Stalder, F. (2006). On the differences between open source and open culture. In *Media Mutandis: a NODE.London Reader*.
- Stallman, R. (1999). The free universal encyclopedia and learning resource. FSF.
- Stallman, R. (2005a). Bill Gates and other communists. *CNET News.com*.
- Stallman, R. (2005b). Historical draft on digital encyclopedias. *Private correspondence*.
- Stallman, R. (2005c). Pigdog Interviews Richard Stallman. *GNUisance*.
- Stallman, R. (2005d). Wale's explanation of the Stallman incident. wikipedia-l.
- Stallman, R. M. (2005e). The Free Software definition. GNU.
- Stallman, R. M. (2005f). The GNU Project. GNU.
- Sterling, B. (2002). *A contrarian view of Open Source (August 3)*.
- Stockwell, F. (2001). *A history of information storage and retrieval*. Macfarlane, Jefferson, North Carolina.
- Suarez, A. and Ruiz, J. (2005). The Spanish fork of Wikipedia. In *Proceedings of Wikimania*.
- Sunstein, C. (2006). *Infotopia: how many minds produce knowledge*. Oxford University Press, Oxford.
- Sunstein, C. R. (2003). *Why societies need dissent*. Harvard University Press, Cambridge.
- Surowiecki, J. (2004). *The wisdom of crowds*. Doubleday, USA.
- Swartz, A. (2006). Who writes Wikipedia? *Aaron Swartz's Raw Thought*.
- Tapscott, D. and Williams, A. D. (2006). *Wikinomics: how mass collaboration changes everything*. portfolio, New York.
- Tavris, C. and Aronson, E. (2007). *Mistakes were made but not by me: why we justify foolish beliefs, bad decisions, and hurtful acts*. Harcourt, New York.
- Thieme, C. (2007). Incivility, re: psychosis. wikien-l.

- Thomas, G. (1992). *A position to command respect: women and the Eleventh Britannica*. The Scarecrow Press, Metuchen, New Jersey.
- Tivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1):35–57.
- Tobias, D. R. (2006). I Am the Very Model of a Modern Wikipedian. wikien-l.
- van Kalken, W. (2006). Re: Secret admin list. wikipedia-l.
- van Maanen, J. (1988). *Tales of the field: on writing ethnography*. University of Chicago Press, Chicago.
- Vesperman, J. and Richardson, D. (2002). *Frequently Asked Questions*. 1.0 edition.
- Viegas, F. B., Wattenberg, M., , and Mckeon, M. M. (2007). The hidden order of Wikipedia. *Proceedings of HCII 2007*, pages 445–454.
- von Hippel, E. and von Krogh, G. (2003). Open source software, the 'private-collective' innovation model. *Issues For Organization Science*, 14(2):209.
- von Krogh, G. (1998). Care in knowledge creation. *California Management Review*, 40(3).
- Voss, J. (2005). Measuring Wikipedia. In *Proceedings of the ISSI 2005 Conference*.
- Vroom, V. H. and Yetton, P. W. (1973). *Leadership and decision-making*. University Of Pittsburgh Press, Pittsburgh.
- Wagstaff, J. (2005). *The New Cliche: "It's the Wikipedia of..."*. loose wire.
- Waguespack, D. and Fleming, L. (2004). *Penguins, camels, and other birds of the feather: the emergence of leaders in the Open Innovation Communities*.
- Waguespack, D. and Fleming, L. (2005). *Penguins, camels, and other birds of a feather: brokerage, boundary spanning, and leadership in Open Innovation Communities*.
- Wales, J. (2000a). Hi... nupedia-l.
- Wales, J. (2000b). Re: bias. nupedia-l.
- Wales, J. (2001a). A culture of co-operation. wikipedia-l.
- Wales, J. (2001b). Re: A proposal for the new software. wikipedia-l.

- Wales, J. (2001c). Statement of principles. Nostalgia Wikipedia.
- Wales, J. (2001d). Why I oppose a cabal. Nostalgia Wikipedia.
- Wales, J. (2001e). Will The Real Nupedia Please Stand Up? *Slashdot*.
- Wales, J. (2004). Re: Request for new Wikipedia: Friulian. wikipedia-l.
- Wales, J. (2005a). Re: Ant: Serbo-Croatian Wikipedia - a policy question? wikipedia-l.
- Wales, J. (2005b). Re: Neo-nazis to attack wikipedia. wikien-l.
- Wales, J. (2005c). Re: Re: A neo-nazi wikipedia. wikien-l.
- Wales, J. (2005d). Re: Re: Static HTML dumps. wikipedia-l.
- Wales, J. (2005e). Re: Sanger's memoirs. Wikipedia.
- Wales, J. (2005f). Re: Status of Wikimedia. wikipedia-l.
- Wales, J. (2005g). Re: [wikien-l] Re: A neo-nazi wikipedia. wikien-l.
- Wales, J. (2005h). Re: Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- Wales, J. (2005i). Re: Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- Wales, J. (2005j). Re: Wikipedia, Emergence, and The Wisdom of Crowds. wikipedia-l.
- Wales, J. (2005k). Wikipedia and defamation: man apologizes after fake Wikipedia post. Air-l.
- Wales, J. (2006a). On "digital Maoism". *Edge*.
- Wales, J. (2006b). Re: Nazi userboxes. wikien-l.
- Wales, J. (2006c). Re: New York Times article. wikien-l.
- Wales, J. (2006d). Re: "should not be written by an interested party". wikien-l.
- Wales, J. (2006e). Re: To: Jimmy Wales - Admin-driven death of Wikipedia. wikien-l.
- Wales, J. (2006f). Wikien-l] 2 Questions for Jimbo about WP:OFFICE and things that aren't labeled WP:OFFICE. wikien-l.

- Wales, J. (2006g). Wikipedia talk:neutral point of view. Wikimedia.
- Wales, J. (2007a). naturally occurring conflicts? Air-l.
- Wales, J. (2007b). Re: Harassment sites. wikien-l.
- Wales, J. (2007c). *Wikimania 2007 interview with Joseph Reagle*.
- Wall, L. (1999). Diligence, patience, and humility. In *Open Sources: Voices from the Open Source Revolution*. O'Reilly, 1 edition.
- Wallace, D. F. (2001). Tense present: democracy, English, and the wars over usage. *Harper's Magazine*.
- Walter (2007). *The English language edition of Wikizine*.
- Walther, J. B. (1996). Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23:1–43.
- Walther, J. B., Gay, G., and Hancock, J. T. (2005). How do communication and technology researchers study the Internet? *Journal of Communication*.
- Weber, M. (1978). *Economy and society: an outline of interpretive sociology (1914)*, volume 1. University of California Press, Berkeley.
- Weber, S. (2004). *The success of open source*. Harvard University Press, Cambridge.
- Webmaster (2007). KDE Women Homepage - Contact. KDE.
- Weinberger, D. (2007). *Everything is miscellaneous: the power of the new digital disorder*. Times Books, New York.
- Wells, H. (1933). The open conspiracy: blue prints for a world revolution (1928). In *The Open Conspiracy and Other Writings*. Waterlow & Sons, London.
- Wells, H. (1934). *Experiment in autobiography: discoveries and conclusions of a very ordinary brain (since 1866)*. The Macmillan Company, New York.
- Wells, H. (1936). The Idea of a World Encyclopedia. *Nature*, 138:917–924.
- Wells, H. (1938). *World Brain*. Methuen, London.
- Wells, H. (2004). A modern utopia (1905). (6424).

- Wells, H. G., Huxley, J. S., and Wells, G. P. (1931). *The science of life: a summary of contemporary knowledge about life and its possibilities I,II,III*. Doubleday, Doran, Garden City, NY.
- Wells, J. M. (1968). *The circle of knowledge: An exhibition to commemorate the 200th anniversary of the Encyclopaedia Britannica*. Newberry Library, Chicago.
- Wenger, E. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge University Press, Cambridge.
- Wheeler, D. A. (2005). *Why Open Source Software/Free Software (OSS/FS)? Look at the numbers!*
- Wiegand, W. A. (1996). *A biography of Melvil Dewey: irrepressible reformer*. American Library Association, Chicago.
- Wikiality (2006). *Main Page*.
- WikiChix (2007). *WikiChix*. WikiChix.
- Wikimedia (2006a). Category:humor songs and poems. Wikimedia.
- Wikimedia (2006b). Don't be dense. Wikimedia.
- Wikimedia (2006c). Elections for the Board of Trustees of the Wikimedia Foundation, 2006/en. Wikimedia.
- Wikimedia (2006d). Polls are evil. Wikimedia.
- Wikimedia (2007a). Conflicting Wikipedia philosophies. Wikimedia.
- Wikimedia (2007b). Steward policies. Wikimedia.
- Wikimedia (2007c). Stewards. Wikimedia.
- Wikipedia (2002). Wikipedia:staying cool when the editing gets hot. Wikimedia.
- Wikipedia (2004a). Wikipedia:assume good faith. Wikimedia.
- Wikipedia (2004b). Wikipedia:etiquette. Wikimedia.
- Wikipedia (2004c). Wikipedia:neutral point of view. Wikimedia.
- Wikipedia (2005a). Wikipedia talk:assume good faith. Wikimedia.
- Wikipedia (2005b). *Wikipedia:Administrators*.

Wikipedia (2006a). AWWDMBJAWGCAWAIFDSPBATDMTD. Wikimedia.

Wikipedia (2006b). Benevolent Dictator for Life. Wikipedia.

Wikipedia (2006c). History of Wikipedia. Wikimedia.

Wikipedia (2006d). Interpedia. Wikimedia.

Wikipedia (2006e). Main page. Wikimedia.

Wikipedia (2006f). Meta:benevolent dictator. Wikipedia.

Wikipedia (2006g). Meta:talk:benevolent dictator. Wikipedia.

Wikipedia (2006h). Nupedia. Wikimedia.

Wikipedia (2006i). User talk:susurrus. Wikimedia.

Wikipedia (2006j). User:raul654/raul's laws. Wikimedia.

Wikipedia (2006k). User:seth Ilys/Dot Project. Wikimedia.

Wikipedia (2006l). Wikipedia:administrators. Wikimedia.

Wikipedia (2006m). Wikipedia:assume good faith. Wikimedia.

Wikipedia (2006n). Wikipedia:assume the assumption of good faith. Wikimedia.

Wikipedia (2006o). Wikipedia:blocking policy. Wikimedia.

Wikipedia (2006p). Wikipedia:blocking policy proposal. Wikimedia.

Wikipedia (2006q). Wikipedia:civility. Wikimedia.

Wikipedia (2006r). Wikipedia:community Portal. Wikimedia.

Wikipedia (2006s). Wikipedia:deceased Wikipedians. Wikimedia.

Wikipedia (2006t). Wikipedia:don't disrupt Wikipedia to illustrate a point. Wikimedia.

Wikipedia (2006u). Wikipedia:editcountitis. Wikimedia.

Wikipedia (2006v). Wikipedia:etiquette. Wikimedia.

Wikipedia (2006w). Wikipedia:five pillars. Wikimedia.

Wikipedia (2006x). Wikipedia:neutral point of view. Wikimedia.

Wikipedia (2006y). Wikipedia:no original research. Wikimedia.

Wikipedia (2006z). Wikipedia:notability. Wikimedia.

Wikipedia (2006aa). Wikipedia:office Actions. Wikimedia.

Wikipedia (2006ab). Wikipedia:please do not bite the newcomers. Wikimedia.

Wikipedia (2006ac). Wikipedia:policy trifecta. Wikimedia.

Wikipedia (2006ad). Wikipedia:schools' FAQ. Wikimedia.

Wikipedia (2006ae). Wikipedia:simple English Wikipedia. Wikimedia.

Wikipedia (2006af). Wikipedia:verifiability. Wikimedia.

Wikipedia (2006ag). Wikipedia:what Wikipedia is not. Wikimedia.

Wikipedia (2006ah). Wikipedia:wikilove. Wikimedia.

Wikipedia (2006ai). Wikipedia:wikipedia is not therapy. Wikimedia.

Wikipedia (2006aj). Wikipedia:wikipediaholic. Wikimedia.

Wikipedia (2006ak). Wikipedia:writing for the enemy. Wikimedia.

Wikipedia (2007). Blocking of Wikipedia in mainland China. Wikimedia.

Wikipedia (2007ab). Charles Van Doren. Wikimedia.

Wikipedia (2007ac). Congressional staffer edits to Wikipedia. Wikimedia.

Wikipedia (2007ad). Criticism of Wikipedia. Wikimedia.

Wikipedia (2007ae). Every time you masturbate... God kills a kitten. Wikimedia.

Wikipedia (2007af). Good faith. Wikimedia.

Wikipedia (2007ag). Help:checkuser. Wikimedia.

Wikipedia (2007ah). Help:template. Wikimedia.

Wikipedia (2007ai). Image:penny Arcade comic-20051216h.jpg. Wikimedia.

Wikipedia (2007aj). List of Buffy the Vampire Slayer episodes. Wikimedia.

Wikipedia (2007ak). Portland Pattern Repository. Wikimedia.

Wikipedia (2007al). Seigenthaler controversy. Wikimedia.

Wikipedia (2007am). Talk:evolution. Wikimedia.

Wikipedia (2007an). Talk:evolution/archive 2. Wikimedia.

Wikipedia (2007ao). User:graft. Wikimedia.

Wikipedia (2007ap). User:jaranda/wikipedia's first IRC chat. Wikimedia.

Wikipedia (2007aq). User:salva31. Wikimedia.

Wikipedia (2007ar). User:the Cunctator. Wikimedia.

Wikipedia (2007as). Web 2.0. Wikimedia.

Wikipedia (2007at). Wikipedia talk:assume good faith. Wikimedia.

Wikipedia (2007au). Wikipedia:10 things you did not know about Wikipedia.
Wikimedia.

Wikipedia (2007av). Wikipedia:about. Wikimedia.

Wikipedia (2007aw). Wikipedia:articles for deletion. Wikimedia.

Wikipedia (2007ax). Wikipedia:assume bad faith. Wikimedia.

Wikipedia (2007ay). Wikipedia:assume stupidity. Wikimedia.

Wikipedia (2007az). Wikipedia:attribution. Wikimedia.

Wikipedia (2007ba). Wikipedia:avoid instruction creep. Wikimedia.

Wikipedia (2007bb). Wikipedia:banning policy. Wikimedia.

Wikipedia (2007bc). Wikipedia:be bold. Wikimedia.

Wikipedia (2007bd). Wikipedia:content forking. Wikimedia.

Wikipedia (2007be). Wikipedia:five pillars. Wikimedia.

Wikipedia (2007bf). Wikipedia:list of administrators. Wikimedia.

Wikipedia (2007bg). Wikipedia:list of shortcuts/project shortcuts. Wikimedia.

Wikipedia (2007bh). Wikipedia:meta. Wikimedia.

Wikipedia (2007bi). Wikipedia:million pool. Wikimedia.

- Wikipedia (2007bj). Wikipedia:oversight. Wikimedia.
- Wikipedia (2007bk). Wikipedia:policies and guidelines. Wikimedia.
- Wikipedia (2007bl). Wikipedia:project namespace. Wikimedia.
- Wikipedia (2007bm). Wikipedia:red link. Wikimedia.
- Wikipedia (2007bn). Wikipedia:replies to common objections. Wikimedia.
- Wikipedia (2007bo). Wikipedia:wikilawyering. Wikimedia.
- Wikipedia (2007bp). Wikipedia:wikipedia Signpost/2007-01-15/2006 in review. Wikimedia.
- Wikipedia (2007bq). Wikipedia:wikiproject. Wikimedia.
- Wikipedia (2007br). Wikiscanner. Wikimedia.
- Williams, G. (2006). *Wikimedia Commons*.
- Williams, R. (1983). *Keywords: a vocabulary of culture and society*. Oxford University Press, New York.
- Williams, S. (2002). *Free as in freedom: Richard Stallman's crusade for free software*. O'Reilly.
- Williamson, M. (2005). Re: Wikimedia Foundation Internal Radio System. wikipedia-l.
- Wilson, D. and Reynard, A. M. (1994). *Interpedia frequently asked questions and answers*.
- Wilson, F. (2003). Can compute, won't compute: women's participation in the culture of computing. *New Technology, Work and Employment*, 18(2):127–142.
- Winchester, S. (1998). *The professor and the madman*. HarperCollins, New York.
- Winchester, S. (2003). *The meaning of everything: the story of the Oxford English Dictionary*. Oxford University Press, Oxford.
- Winner, L. (1986). Do artifacts have politics? In *The Whale and the Reactor*, pages 18–39. The University of Chicago Press, Chicago.
- Woolgar, S. (1991). The turn to technology in social studies of science. *Science, Technology, & Human Values*, 16(1):20–50.

- Yankelovich, D. (2001). The Magic of Dialogue. *The Nonprofit Quarterly*, 8(3).
- Yeo, R. R. (2001). *Encyclopedic visions: scientific dictionaries and enlightenment culture*. Cambridge University Press, Cambridge.
- Yoo, Y. and Alavi, M. (2004). Emergent leadership in virtual teams: what do emergent leaders do? *Information and Organization*, 14(1):27–58.
- Yukl, G. A. (1981). *Leadership in organizations*. Prentice-Hall, Englewood Cliffs, New Jersey, 1 edition.
- Zakon, R. H. (2005). *Hobbes' Internet timeline v8.1*.
- ZenZagg (2004). *The whole is greater than the sum of its parts*. everything2.