San Jose State University

SJSU ScholarWorks

2009

AMCIS 2008 Panel Report: Aging Content on the Web: Issues, Implications, and Potential Research Opportunities

Brent Furneaux
York University

Timothy R. Hill
San Jose State University, timothy.hill@sjsu.edu

Wayne Smith
California State University, Northridge

Shailaja Venkatsubramanyan
San Jose State University, shailaja.venkatsubramanyan@sjsu.edu

Jingguo Wang
University of Texas at Arlington

Follow this and additional works at: https://scholarworks.sjsu.edu/faculty_rsca

1-2009

AMCIS 2008 Panel Report: Aging Content on the Web: Issues, Implications, and Potential Research Opportunities

Brent Furneaux
York University

Timothy R. Hill
San Jose State University

Wayne Smith
California State University, Northridge

Shailaja Venkatsubramanyan
San Jose State University

Jingguo Wang
University of Texas at Arlington

See next page for additional authors

AMCIS 2008 Panel Report: Aging Content on the Web: Issues, Implications, and Potential Research Opportunities

Authors

Brent Furneaux, Timothy R. Hill, Wayne Smith, Shailaja Venkatsubramanyan, Jingguo Wang, Anne Washington, and Paul Witman

Communications of the Association for Information Systems

CAIS

AMCIS 2008 Panel Report: Aging Content on the Web: Issues, Implications, and Potential Research Opportunities

Brent Furneaux
York University

Timothy R. Hill
San Jose State University

Wayne Smith
California State University, Northridge

Shailaja Venkatsubramanyan

San Jose State University

Jingguo Wang
University of Texas at Arlington

Anne Washington
George Washington University (GWU)

Paul Witman
California Lutheran University

Abstract:

Since its inception in the early 1990s, the World Wide Web (Web) has grown enormously. According to the "official Google blog" (Google 2008), the Web had 1 trillion (as in 1,000,000,000,000) unique coexisting URL's as of July 25, 2008. Given the exponential growth of the Web over time, an issue that is likely to gain prominence is that of outdated information. This is especially important to study since many of us rely on the Web to find facts in order to take decisions. For example, for students and researchers, the "date" of a document is important for scholarship and student work. However, getting an accurate date on content is challenging, and furthermore, outdated pages that are not deleted from Web servers will continue to be returned in response to Web searches. The panel, held at the 2008 Americas Conference on Information Systems in Toronto, Canada, identified a number of research issues and opportunities that arise as a result of this phenomenon.

Communications of the Association for Information Systems

I. INTRODUCTION

Given the size of the Web, its decentralized nature, the inability to figure out the date of publication of many Web documents, and the lack of an existence of an expiry date on Web pages [Maurer 1998], the existence of outdated information on the Web is both inevitable and problematic. Such content may serve no purpose and may even cause harm. Another feature of the Web that contributes to this issue is that a lot of data is duplicated. Of all the text documents on the Web, about 46 percent of the documents have at least one near-duplicate document which is identical except for low-level details, such as formatting. In fact, 5 percent of text documents on the Web have between 10 and 100 replicas [Shivakumar and Garcia-Molina 1998]. When duplicated information becomes outdated, mistakes are proliferated, diminishing search efficiency, an issue for public content (such as blogs and sites) as well as private content (such as e-mail and financial accounts).

The flip side of the issue is the danger of valuable content disappearing with age. Pitkow [1998a, 1998b] reported that around 5 percent to 8 percent of requested hyperlinks on the Web are broken (i.e., returned an error). Lawrence et al. [2001] found that many URL citations in research articles become invalid as early as a year or two after publication. This problem is being addressed by the Internet Archive project [http://www.archive.org/index.php], working to prevent the Internet and other "born-digital" materials from disappearing into the past, though many challenges remain. Still, the problem of retiring outdated content is more challenging and has not received much research attention toward a solution to date.

A study [Cho and Garcia-Molina 2000] conducted over a four-month period from February 1999 through June 1999 found that more than 20 percent of 720,000 documents tracked changed daily, and 70 percent of the documents changed to some degree within four months. More than 70 percent of the documents remained visible (regardless of changes of content) for more than one month. Documents in the .com domain were found to be updated more frequently than average – 40 percent changed every day, and 50 percent changed within 11 days (as opposed to 50 days for half of documents across all domains to change). The study also concluded that the average lifespan of commercial documents is less than the average lifespan of .edu and .gov documents. When content that changes frequently is replicated or there are links that have been created to that content, it is very likely that the replications will be outdated, and the links broken. When content does not change for long periods of time, it could be an indication that it is not being maintained and updated over time.

Outdated information can have many consequences. For instance, the existence of outdated and incorrect medical, pharmaceutical, and legal information can lead to physical and monetary losses. There have been reports that there is a significant amount of wrong and potentially dangerous medical information on the Web. [ABC News.com 2001; University of Michigan News Archive 2002; Norton Article Library Sep. 2007]. A book about bankruptcy mentions that some of the information on the Web may be of questionable accuracy since the information may be too old even if it was completely error free when it was initially posted [Jurinski 2003].

A panel presented at the 2008 Americas Conference on Information Systems (AMCIS), held in Toronto, Canada, in August 2008, brought together different perspectives on aging content on the Web, and the discussion surfaced several core issue themes and driving forces:

- **Definition of the term "outdated"**—The size of the Web has been a matter of debate. Estimating the proportion of the Web that is not current would be a greater challenge. To figure out how much of the Web is outdated, one would have to define the criteria to determine if a document is outdated, which may involve determining the relationship between dormancy, utility, and age. Algorithms may be needed to organize documents along a timeline using semantic and other heuristics.

- **Managing demise of outdated information**—Will current existing forces such as Web server account expiry and community editorship (such as in Wikipedia) cause the natural demise of outdated information? Will the Web naturally correct itself without any human intervention?

- **Aging and Web Standards**—There is no central authority or body that certifies Web data as being current or accurate. Standards age too and constrain the pragmatic boundaries of implementation. What are the standards and legal issues that impact the ability to address aging Web content?

- **Impact on Information Search**—Search engines could selectively omit indexing what they consider to be outdated information by altering their algorithms. What is the impact of content aging on strategic decision making?
- **Library versus the Web**—How do librarians keep information up to date? Can any of those techniques be extrapolated to the Web? Are there any problems unique to the Web?

For some of these issues, there is existing literature that supports their emergence, while others were recognized as needing seminal work to be fully articulated. In all cases, a rich array of research opportunities presents itself. The remainder of this paper elaborates on each of the previous issues and suggests compelling research questions and promising directions. Section II addresses the definition of "outdated" in this context, while Section III discusses managing the demise of outdated content. Section IV considers aging and Web standards. Section V examines the impact of content aging on Web search, and Section VI explores implications to be drawn from the analogy to content age management in the library. We conclude the paper with a summary, Section VII.

II. DEFINING "OUTDATED"

The panel discussion crystallized perhaps the most essential question: What does it mean for Web content to be outdated or obsolete? Although obsolescence has been characterized in the information sciences literature as a decline in user perceptions of the usefulness or validity of information [Alvarez et al. 2000; Egghe and Rousseau 2000], we noted that there is significant room for interpretation in any determination of obsolescence. For example, research examining obsolescence of the traditional academic literature has frequently been based on the premise that falling citation levels are indicative of both aging and obsolescence [Alvarez et al. 2000; Line 1993]. This pre-Web perspective tends, however, to overlook the important distinction between information use and information usefulness. Some information may be extremely useful on infrequent occasions–low levels of information use can only be considered *suggestive* of obsolescence. Furthermore, in the age of the Web, rampant duplication makes citation tracking problematic, particularly outside the academic literature arena where ethical pressures and accountability may be less influential. Granularity complicates the issue as a single page on the Web may combine an array of informational components with differing levels of obsolescence and update rates. Additionally, it was noted that numerous economic counter-incentives work against the ideal of a fully current Web, or even a fully date-determinable corpus since, for example, sites marketing content are disinclined to even reveal the origin dates in order to maximize the valuable life of their material. Despite such challenges and the presence of some degree of ambiguity, the understanding of information obsolescence provided by the information sciences literature serves to underscore some of the thoughts surrounding obsolescence that were raised during our discussion.

The definition of obsolescence offered by information scientists suggests that it is multiply determined, being triggered by such things as new innovation, improved understanding, and the detection of flaws or shortcomings that significantly undermine information validity [Egghe and Rousseau 2000]. Our discussion explored a number of possible factors that might contribute to the obsolescence of Web content. Among these were such broad and relatively ambiguous events as scientific discovery and changes in culture and taste, as well as considerably less ambiguous events such as the death of an individual who has been posting personal content on the Web. Reflecting on the impact of scientific discovery on obsolescence, it was observed that the process by which these discoveries render established knowledge obsolete can be both slow and uncertain. Discussion surrounding the implications that the death of an individual might have for his or her personal Web content highlighted some of the difficulty associated with establishing obsolescence. Although the continued availability of such content might be discomforting for some, others might see it as an opportunity to celebrate the life of a loved one. There is a growing recognition that, as outdated information can retain value past its otherwise useful life, the ideal is not pure currency but rather *presentation in temporally-appropriate context so that content may be applied in uses suitable to its age.* "Expiration dates" are not simple matters then, but rather are seen as being dependent on application.

It was clear from our discussion that assessments of obsolescence are heavily dependent on judgments by individuals or groups of individuals and that it must therefore be considered a socially constructed phenomenon [Brookes 1970; Kuhn 1962]. As a result, determining the obsolescence of Web content might necessitate the identification of who is drawing upon this content and a development of some understanding for the nature of their interests [Brookes 1970; Nicholas et al. 2005]. The possibility that Web content could be moving toward obsolescence for one group of stakeholders while simultaneously becoming more important to another group was highlighted by an observation that the Web serves as both an information store and an archive. It was, for instance, noted that information considered obsolete by scientists might have historical value to at least a small group of individuals. Attention has also been drawn to this possibility by research examining the obsolescence of scientific literature. This research suggests that two processes of obsolescence may be at work [Brookes 1970]. The first of these processes is a relatively rapid decline in scientific interest that begins to occur within a few years of publication, while the second is a much slower decline in historical interest [Brookes 1970].

Much of the research on the obsolescence of the scientific literature was motivated by the need to address the problem of space limitations surrounding the storage of physical volumes [e.g. Brookes 1970]. Virtualization of information on the Web and dramatic reductions in the cost of data storage may, therefore, increase the extent to which information is retained for historical or archival purposes. Research in the domain of information lifecycle management suggests, however, that the notion of obsolescence remains important as organizations struggle to minimize the costs associated with the storage of vast quantities of information [Tallon and Scannell 2007].

Rather than being either useful or not useful and valid or not valid, information can exhibit varying degrees of both usefulness and validity and therefore be characterized as exhibiting varying degrees of obsolescence. Building on this understanding, we recognized that there is variability in the rate of obsolescence of the various "pieces" of information contained in even a single journal article or on a single Web page. Although the potential for differential rates of component obsolescence has been recognized in relation to physical products [Sandborn et al. 2007], much of the research related to the obsolescence of the scientific literature has tended to view whole journal issues or volumes as discrete bundles of information that uniformly suffer obsolescence. Unpacking information in the way that we suggest does, however, introduce a considerable measure of complexity to efforts to understand and address obsolescence.

Although limited discussion covered the rate at which Web content might become obsolete, it is clear that this rate can vary considerably, depending on the factors driving obsolescence in any given context. Research suggests, for instance, that scientific publications intended to support current research age quite quickly in comparison to those used for archival purposes [Egghe and Rousseau 2000]. Obsolescence rates can also be heightened in subject areas characterized by rapid technological innovation and attenuated when the subject matter is descriptive, conceptual, or critical in nature [Line 1993].

Our discussion of the notion of obsolescence thus covered a broad range of conceptual issues and concerns. The literature appears to support some of our key ideas, though a review of this literature suggests that there is considerable opportunity for further exploration of information obsolescence. It should be noted that the importance of obsolescence is likely to vary depending on the risks associated with relying on the wrong information, and research attention may therefore be particularly warranted in high-risk contexts. Examples of research questions include the following:

- What is a substantive definition for what it means to be outdated, and what are effective age metrics?
- What document standards are needed to label documents with context-appropriate expiration dates?
- How can content presentation be designed so that the user can readily choose to look at the content but still be made aware of its status, i.e., how can temporal context be conveyed effectively through the user interface? This could be akin to the types of work done to highlight security status in the browser, at least in some situations, and might be envisioned as being developed as a browser helper object or add-in.
- What kinds of client- or server-side tools are needed to identify and properly respond to outdated content?

Reference disciplines include Digital Preservation and User Interface Design [Dhamija and Tygar 2005; Jakobsson, et al. 2008; Thibodeau 2002], specifically around awareness of the "freshness" status of a page, much like the need to make users aware of the security status of a particular page.

III. MANAGING DEMISE OF OUTDATED INFORMATION

Having discussed the various issues surrounding the definition of "outdated" for the Web, the panel then discussed the potential mechanisms for recognizing outdated content. These include document dormancy, and the inclusion of revision and expiration dates in the document or in the document metadata. Revision and expiration dates in the document may assist human readers or page parsers in detecting outdated content. These dates in the document metadata could be used for automated detection of potentially outdated content.

Document dormancy (based on the date of last document revision taken from the document's properties) provides some indication of the freshness of the document. For a search engine, it is also possible to compare the current version with the most recently indexed version to determine if the content has changed since the last indexing. However, without knowing the context of the document, the date of last update really provides relatively little information. For example, meeting notes, once published, are unlikely to be amended or updated, yet they are still a true and correct representation of a past event.

Other indicators of potentially outdated content include logical issues. Broken links in a document may reflect a lack of ongoing maintenance, but they may also be merely a symptom of temporary problems at the linked sites. The

content may be incompatible with current versions of software or may have been created for software that is no longer available.

Numerous tools exist that may contribute to automatically detecting outdated content, flagging it as such, and perhaps even depreciating its value and impact in a way appropriate to its age, considering the intended use. Most put the burden for marking and determining content freshness on the content author or the server. Most current content management tools provide mechanisms for specifying expiration dates at which content will be removed from public display (though it may still exist on the server). Van Harmelen and van der Meer proposed a tool called WebMaster that used semantic content elements in Web documents, including revision dates, to allow users and intelligent clients to evaluate the freshness of documents [van Harmelen 1999].

The requirement for authors to take specific action to cause their work to be tagged for freshness, and thus to be later automatically removed, seems likely to be useful only to a relatively small portion of Web authors, and will likely be viewed as a burden or ignored by others. If this capability were built into high-usage authoring tools, then the metadata would be automatically updated without specific author intervention, and the usage of this metadata might reach critical mass. It is also likely that if this metadata significantly affected search engine results, then a further natural outcome would unfortunately be the modification of the metadata to improve search engine results.

One might reasonably ask whether outdated content could be maintained "naturally." Such a demise of outdated content might happen if content was reliably maintained. A significant portion of Web content is created by individual bloggers and personal Web pages. Statistics for three major social networking sites (MySpace, Facebook, and Reunion) show more than 200 million combined active users [Owyang 2008], each of whom may have multiple pages. This does not count inactive users nor does it count other major social networks like BeBo. An aggregate estimate of blog activity indicates over 184 million blogs worldwide, as of early 2008 [Technocrati 2008]. Clearly, many of these may be inactive or have not produced content that will draw large readership. However, they still contribute to the total volume of Web content and to the potential for outdated content.

An interesting point of discussion in the panel is the issue of the demise of the content author. A convenience sample of a variety of Web hosting and blogging sites was surveyed to find their policies for password retrieval in the event of the death of an account owner, so that a survivor could take over ownership and operation if so desired. The survey was conducted in October 2007, and included approximately 12 major Web hosting and blog hosting services. In that survey, none of those services had publicly posted policies for dealing with the death of an account owner. Indeed, in many cases (because the service was ad-supported and therefore not charged to the user), the services often have no real identification information for their users. As such, password recovery can really only be performed if another individual knows the answers to the secret questions.

The death of an account owner is obviously an extreme case, but it is a real one and is indicative of the types of pragmatic issues that arise with content that is individually owned and maintained. Should the owner, for any reason, choose to cease making updates, the content will, in many cases, remain visible for an indeterminate period. For ad-supported or otherwise "free" hosting sites, the site earns revenue for content hits and is thus unlikely to remove that content unless the cost of hosting exceeds the revenue earned, or unless they have explicit account expiration policies. Paid accounts, on the other hand, are more likely to have their content removed once payment for the site has stopped. In this case, when lack of updates and lack of payments go together, it is likely that the now-stale content will more quickly be removed from public view. That content could well survive for some time in search engine caches, of course, as well as archives and other mirroring facilities.

Other document approaches, like Wikis (e.g., Wikipedia), employ groups of authors and editors for content maintenance. In this type of situation, unless a topic is so specialized that it draws only one contributor, there is a greater likelihood that the community will continue to maintain content even if one of the primary authors no longer continues.

One example of the need for explicit management of outdated content comes from medical documentation. Researchers [Shekelle et al. 2001] assessed the rate at which medical guideline documentation became out of date. Of a set of 17 guidelines, the researchers found that approximately half of the documentation became outdated in 5.8 years. In this case, the authors declared as "outdated" any material that was no longer correct, due to newer findings in medical research, or due to changing protocols for medical treatments, based on peer-reviewed published evidence that caused major or minor updates to be required for the guidelines. Outdated documents may provide information that is no longer valid and may thus increase the risk of incorrect treatments or approaches to medical issues.

As part of a strategy for ensuring that medical documentation is clearly identifiable as to its currency, the United States National Library of Medicine (NLM) specifies permanence levels for its documentation that is published on the Web [National Library of Medicine 2007]. There are four permanence levels, including:

1. *Unchanging content*—committed to be permanently available, and content will not change (e.g., meeting minutes)
2. *Stable content*—committed to be permanently available, but contents may be subject to minor revisions (e.g., annual reports)
3. *Dynamic content*—committed to be permanently available, but content subject to change
4. *Non-guaranteed content*—no commitment for availability, and locator information may change as well (e.g., FAQs).

These permanence levels are then used to ensure that documentation is clearly identified to users, as to its intended permanence, and as to its currency. Markers of currency include not only the permanence level but also the dates of publication and last revision and the date for the next planned review of the document. Clearly, this model requires considerable discipline to execute, along with a metadata recording system that can assist in managing the process over a large collection of documents.

Literature reviews indicate that managing the demise of outdated information has received relatively limited attention despite the considerable social, political, and technological implications. It offers rich research opportunities that lie at the interface between people and technology, widely regarded as a defining characteristic of the IS discipline. Some examples of specific research questions raised by the panel include:

- What drives the obsolescence of Web content?
- What are the processes by which obsolescence occurs?
- How does the structure of the Web and the technologies that are used to access it impact obsolescence?
- How can technology be used to facilitate the identification and removal of obsolete content?
- How and to what extent do users establish the obsolescence of Web content?
- How are the interests of some stakeholders served by the presence of obsolete content?

Addressing such questions will be challenging and will draw upon reference disciplines including information sciences, information management, organizational theory/organizational behavior, sociology, psychology, and philosophy (ontology and epistemology). Though challenging, the effort is seen to be a worthy investment, given the impact and import of the results.

IV. AGING AND WEB STANDARDS

The panel also considered content aging and its relation to the accessibility of information. In his popular book, *The Sciences of the Artificial*, [Simon 1981] Simon reminded us that everything we create in an artificial world is ultimately shaped by the external environment. To the extent that IS researchers make the assumption that the Web is indeed an information technology artifact that represents some fraction of empirical reality [Bray, et al. 2006], then an understanding of the differences and concomitant tensions between the Web and the end-user environment is paramount. The "aging" of content is one such tension. In effect, both researchers and practitioners desire the "aging" of the content to be consonant in time and space with the "aging" of the end-user. Minimally, the aging of the content should not occur before the aging of the end-user population (that is, inadvertently "pre-aged"). Optimally, the diffusion of the standards and protocols of Web technology, including adoption by the automated and manual authoring communities, should be at least as rapid as the maturation of the end-user requirements for accessibility and quality.

Some aspects of Web data and document life-cycle management parallel other substantive technological developments, including traditional libraries and other central, local, in situ repositories; however, some aspects of Web technology diverge from this model, often considerably. With respect to the "aging" of content, there are many aspects of contemporary Web measurement and management. Here we highlight particularly challenging aspects that fall within the regulatory and technological domains of the life-cycle.

Of all the aspects of the Web aggregate system--people, data, networks, software, hardware—one particularly difficult and complex component of the Web has the force of law behind it. This law is known at the federal level as "Section 508" [Section 508 2008]. This law articulates specific technical requirements for Web accessibility by physically challenged (visual, aural, motor, and cognitive) individuals. Broadly, the Section 508 law and specifications are to electronic pathways what the "Americans with Disabilities Act" (ADA) law and specifications are to physical pathways. Much of the Section 508 law is based upon the Web Content Accessibility Guidelines [WCAG

2008], which are technical recommendations managed by the World Wide Web Consortium [W3C 2008]. Ensuring that Web sites and pages are inclusive by design and by default remains non-trivial, even for institutions of higher education in a large, progressive state [Smith 2008]. Inadvertent "pre-aging" can occur when Web pages are inaccessible to specific subgroups of end-user populations.

The case of NFB, et al. versus Target Corp. is illustrative [DRA 2008]. The final settlement of this case occurred on August 27, 2008. This case involved the inability of a blind individual to use the Target Stores (a retailer with both a physical and online presence) Web site for many functions, including basic purchases and account setup for gift cards. The Web site had numerous problems, including the use of graphics as "submit" buttons, non-identified graphics and labels on fill-in forms, and poor navigation hierarchy. This case was eventually escalated to a case-action lawsuit. Among other things, the settlement called for periodic, detailed evaluation of the Web site by an outside consultant, qualitative assessment of Web accessibility by a group of physically-challenged individuals, and an ongoing, collaborative relationship between the NFB and Target Stores in both requirements engineering and Web page design and development for at least three years. This example illustrates the fact that the spirit and letter of various state and federal laws can be specific enough, even in a rapidly evolving technology area such as the Web, to administer justice and promote equity. Providers of goods and services can be held liable for "pre-aging" content; that is, content that is not accessible to specific subgroups of end-user populations.

The main unit of composition on the Web is the Hypertext Markup Language (HTML) page, and the primary access device (at least currently, until mobile devices replace desktop devices) on the Web is the browser, such as Internet Explorer (IE), Mozilla Firefox, or Apple Safari. The specific architectural, engineering, and construction technical details, such as syntax requirements, technical recommendations, and underlying protocols, individually define an HTML Web page and collectively define the Web environment. The assumptions and processing paths programmed into a browser define what a browser will display to a user and how the HTML and related files are to be interpreted. This is the content, or data, of the Web page.

But data is partly managed by metadata that describes the contents of the data. By extension, there is "meta-content" (relatively invisible structure intended for consumption by "machines") for each piece of "content" (relatively visible detail intended for consumption by "humans"). Not only does the W3C manage technical recommendations for content, the W3C also manages the technical recommendations for meta-content. Germane to the idea of content aging, there are recommendations for schemas (XML/Schema), semantics (RDF/OWL), and Privacy (P3P). There are other, key recommendations, including those of Subscription and Aggregation (RSS/ATOM) files. If one or more of these meta-content recommendations are not used in conjunction with the HTML/CSS content recommendations, then the content risks being "pre-aged." Even the built-in <meta> tag in HTML, especially when used with a widely known lexicon such as the Dublin Core [DC 2008], can be used to store and forward dates. The main standard for semantic interoperability, RDF [RDF 2008] has remained stable since 2004. However, just as the distinction between "flat-file" databases and databases in third-normal form is nontrivial, the ability for authors to extend their HTML/CSS (syntax) knowledge to RDF/OWL (semantics) knowledge is non-trivial. This new ability requires both a conceptual leap and a technological tool use leap and may ultimately result in a transition from less manual Web page generation to more automated page generation.

One final issue on the subject of content aging is worth noting. Much of the aging discussion, including the engaging dialogue at the AMCIS panel, assumes that the content is chiefly static text. It is not clear how to "age" dynamic content, including "content" that emerges via a scripting language, such as JavaScript. Not only can parts of a Web page age at different rates, different versions of a scripting language (including the versions deeply embedded in different browsers) may interpret a command file differently. This ability to tightly interweave "dynamic content" with "static content" may define the spirit and letter of "aging" in the future. In effect, there may be no such thing as a "document." The community may conclude that it's all just data, managed by appropriate meta-data, living within the confines of technical recommendations and legal requirements.

Again these issues present rich research opportunities. For example, given existing Web content standards and emerging subscription standards, there are research questions surrounding the diffusion of "blogs," particularly in how they (that is, content or pointers to content organized in some type of chronological structure) will displace or supersede in importance "non-blogs" (that is, "traditional xHTML Web pages") that embed no chronological or other date-related data or meta-data in the "document."

Studying the technology and diffusion of semantic authoring (either by hand or generated quasi-automatically by a programmatic process) is important as well. The most promising technology is the RDFa standard for embedding semantic information ("meta-data") into new or existing XHTML files. Since the Resource Description Framework (RDF) uses a logic chain (subject-property-predicate) at root, using RDF (as a separate file) or RDFa (embedding the meta-data as a micro-format inside the same file used for presentation) provides an ample platform to specify

the granular details of how a document (or container of data) should be aged. A key question is the degree to which a human author can reliably know the date of aging ex ante and encode it properly without coding errors. These questions involve analysis of authoring, consumption, archiving, and ultimately, aging.

Related disciplines include computer science, library science, and human computer interface (HCI) studies. Computer science and informatics paradigms dominate the normative (prescriptive) and informative (descriptive) specifications, particularly within the World Wide Web Consortium (W3C). To the extent that technology and engineering "best practices" regarding content aging, or any other Web application, is measured and managed as a successful strategy, the W3C is likely to be the reference. This is particularly crucial as semantics continues to augment syntax within digital published works. Library science and by extension, practicing librarians, have developed standards, policies, and procedures to deal with issues of content aging (including knowing *when not to specify it*, such as in the inside front cover of a procured book). HCI researchers live at the intersection of several disciplines and interfaces. One area of research is the differing nature of "content aging" required for human consumption and required for machine consumption.

V. IMPACT ON INFORMATION SEARCH

The next topic of discussion in the panel was how content aging affects accessibility of information by search engines. According to the model (see Figure 1) presented by Venkatsubramanyan and Kwan [Venkatsubramanyan and Kwan 2008], one of the distortions in Web search occurs when content retrieved by the search engine is not relevant to the decision problem. This can be: a) content retrieved that is current but not relevant to the query; and b) content retrieved that is not relevant due to inconsistencies between the indexed content and real world phenomena. In a realistic situation, the user will receive less than perfect information as a result of the inconsistencies in the Web search environment.
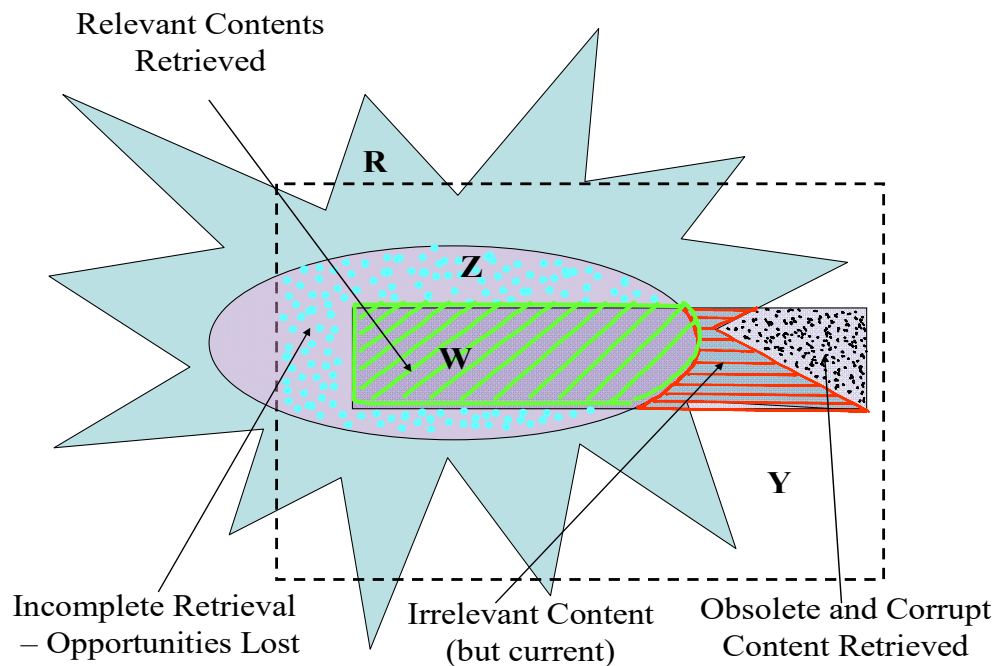

Figure 1. Set Diagram Showing Search Contents and Results (Figure not to scale)

Any decision needs relevant information and nearly all problem solving and decision making behavior relies on information search [Simon 1981]. The way individuals engage with information is being changed by the Internet. With the growth of the Internet and the World Wide Web, nowadays more and more individuals rely on search engines to find the information they need. Recent studies from Pew Internet and ComScore show that about 60 million American adults use search engines on a typical day, and 90 percent of the Internet users use search engines [Rainie and Shermak 2005]. Individuals who feel they have something important at stake are more likely to engage intensely with online search [Fox 2008].

In a perfect world, we would all retrieve high-quality, perfect information every time we do online searches. The most basic requirements for quality information are [e.g., Floridi 1996]:

- *Objectivity*. We expect that the information is presented in a manner free from propaganda or disinformation.
- *Completeness*. We expect that the information is a complete, not a partial, picture of the subject
- *Pluralism*. We expect that all aspects of the information are given, and the information is not restricted to present a particular viewpoint, as in the case of censorship.

However, our world is not perfect. Online searchers confront the question of information accuracy on a daily basis, because anyone can post anything on the Internet. The quality of the information returned by search engines may not be guaranteed. In addition to these three basic requirements, we should also expect the most recent information available for our decisions, instead of obsolete or outdated information. This issue is especially important because content and information over the Internet are updated frequently. Sustersic et al. (2006) investigated 22 Web sites and tracked 14,523 distinct Web resources in seven weeks. Sixty-six percent of these tracked Web resources were HTML files. They found that on average the tracked resources were changed nine times in seven weeks, and 46 percent of resources were changed at least once. For those changed, the mean time between changes was 21 hours.

As end users, we may wonder how search engine or other systems, for example Web caching or content delivery networks (CDN), reflect content changing. Web resources themselves carry some time-related attributes. The most common time-related attributes include DATE, EXPIRES, and LAST-MODIFIED TIME. DATE normally indicates when the resource was created. EXPIRES is the date and time when the content may be obsolete. A lot of Web caching systems and CDN use this attribute to decide whether caches or edge servers need to make a request to the origin server for a fresher copy of the content. LAST-MODIFIED TIME reflects when the resource was last modified. In Web caching systems and CDN-s, there are other complicated synchronization mechanisms, trading off between the freshness of the content in the edge servers and the cost of updating/communication. The details of these mechanisms are beyond the scope of this discussion [e.g., Rabinovich and Spatscheck 2002, for some more discussion on this].

Search engines like Google normally update their index on a regular basis (for example, every four weeks). Until content crawlers check Web content the next time, there is no synchronization mechanism in place that can make sure when and how the changed content will be reflected in the search results. Further, as we have discussed earlier, ranking algorithms like PageRank probably do not favor recent and new content over old content in their ranking. In another words, the age of a piece of content may have positive effects in the ranking. The reason could simply be that the new content has not attracted much attention yet.

Besides content aging, there are also other issues related to the quality of information over the Internet. We may be caught by bogus information, which may appear in some reputable news Web site. We see different types of online deceptions [Rowe 2005], including Web scams, falsified research, hoax sources, and so on. Will a search engine know which source is more reliable or have higher quality? A search engine may be able to determine the quality and reliability of Web sources based on certain assumptions. For example, a reliable source will attract more people to visit, and it may have more other Web sites link to it or "cite" it. The reliable source could also have more outlinks. However, all these metrics cannot guarantee a site is reliable, because these are observed phenomena, and the fraudulent may generate those metrics artificially like click fraud.

Content aging and active deceptions over the Internet may significantly impact individuals' search behaviors. As Google could not tell us explicitly which piece of information is a scam and which piece of information is more recent, it could become a challenging task to identify reliable and quality information, especially for those inexperienced and unknowledgeable in the field that they tend to search. And, in fact, individuals' judgment and decisions are significantly affected by the ranking of the content in the search result list [Joachims 2007]. Recent studies based on the number of queries submitted in individual search sessions showed that individuals tend to search more intensely for healthcare related information than they do in searches related to other matters [Wang et al. 2008]. Maybe one of the reasons is that individuals try to reduce the uncertainty regarding some important information by looking at different sources, since the quality of information in healthcare across different online sources varies [see Eysenbach et al. 2002 for a review].

Content aging and information quality are the emerging issues with the development of Internet and online search industries. More research effort is needed to understand how content aging and information quality over the Internet affect individual information seeking and decision-making. Examples of some research questions include, but are not limited to:

- Are users of search engines aware of the issues of content aging and information quality?
- How do users distinguish the reliability of information source? What are the strategies they use?

- How much do users trust information returned by search engines? How does the rank of a resource in the result list affect users' trust of the source?
- How does users' trust of online sources affect users' search behaviors and strategies?

To address these questions, we may draw from other disciplines.   Literature in decision-making suggests that information is important for any decision, and the availability of the information affects individuals' decision-making behavior. As decision-making needs quality information, studies in information quality look into how we define quality information. Research in marketing as well as in library science studies individuals' behavior related to information seeking (for example, how does an individual search for information, and to what extent will an individual search).

VI. THE LIBRARY ANALOGY

The panel next focused on whether experiences from the analog world could shed light on options for dealing with aging digital information. People have been grappling with long-term maintenance of information and objects for millennia. The management of long-term storage of information objects traditionally has been divided between archives and libraries. The logic behind the separation between libraries and archives may provide insight into handling aging content on the Web.

An archive contains unique items that document everyday activity or functions for a specific group of people [Cox 1994; Hensen 1989; Ritzenthaler 1993]. For instance, the National Archives and Records Association, NARA, documents the activity of government employees and agencies, while the Institute of Jazz Studies documents the lives of jazz musicians. The removal of objects in an archive depends upon its mission. Some archives use retention schedules which stipulate how long to keep certain types of records. Most items that are considered a permanent part of the collection are carefully preserved and cataloged.

A library contains published items that support the knowledge needs of a particular community either geographic or institutional [Buckland 1992; McArthur 1986]. For instance, a corporate library supports its organization while the town public library supports its local citizens. The removal of objects in a library occurs routinely as new information items are available. For instance, a first edition of a textbook may be weeded out when a later edition is published. Traditionally archives, which document an activity in time, weed their unique items rarely, while libraries, which document ideas, change their published items frequently.

The division between archives and libraries is being challenged by the nature of digital objects. Analog items can be distinguished between unique items and published ones. However this division won't work for digital items. The very definition of a digital item is that it is not unique. A digital copy is a perfect replica of the original [Samuelson 1998]. The determination of the object's status based on its intrinsic value cannot be used for digital objects. Another mode of thought is necessary to understand the difference.

The difference between the archival mission of tracking an activity and the library mission of tracking an idea could be useful in a digital realm as a user tries to understanding aging content. Digital materials that track an idea are constantly changing and could be considered outdated. These digital materials capture the most recent thoughts and arguments regarding an idea, for example, a Wikipedia page. Information pages need dates to confirm that the page has been updated. The date is particularly important if, for some reason, the old page is not subsumed by a new one. Digital materials that document an activity in time would never be considered outdated. These digital materials capture a moment in history, for example, Tim Berners Lee's first Web page. Historic activity pages require dates so information could be placed in context. Any page, which contains news, might best be considered an archival record, which documents a current activity.

The recent plunge in stock price for United Airlines is an example of the problem when there is not a distinction between historic and current pages. On September 8, 2008, a Florida investment service mistakenly quoted a six-year-old *Chicago Tribune* article because it appeared at the top of a Google search [Abels 2008]. The false news that United Airlines was going into bankruptcy sent the stock price into a nosedive within minutes. The *Chicago Tribune* article contained links to contemporary pages, however, the information was undated. Without the necessary date, the page appeared to be a current information page when it was actually a historic activity page.

Government information is a good example of the tension between current information and historic activity when dealing with aging content. For example in the case of legal information, citizens may simultaneously want to see the current law as well as any relevant pending legislation. Citizens might turn to the Federal Emergency Management Agency to find the best current information about an ongoing disaster. However, another citizen may be looking for relief information for a disaster that already happened. The government must balance the public expectation for maintaining historic activity with that for publishing current information.

The expectation that older government information remain available began just after the September 11, 2001, attacks, when some open documents became classified. This caused an uproar with citizens worried about political motives and historians worried about future research [Hensen 2001; Nelson 2002]. While the controversy focused on the issues of access to government information, it brought to light the process that the government uses to weed aging content.

While most federal agencies maintain current information Web sites, specific agencies handle historic aging information. In the tradition of archives, these items document ongoing activity. The groundbreaking E-Government Act of 2002 established the dissemination of government information through the Internet. Following the structure of how paper documents are handled, the National Records and Archives Administration (NARA) began to receive electronic records submissions. However, other federal agencies play a role. The Government Printing Office (GPO), the federal government publisher, is responsible for permanent access to the documents it publishes (http://www.gpo.gov). To accomplish this, GPO runs a program that sends paper publications to libraries across the country and also provides access to these documents through the Internet [Jacobs et al. 2005]. The Library of Congress, a national library, has led the way on digital preservation and cataloging technology and has sponsored several projects that harvest and catalog government Web sites (http://www.digitalpreservation.gov).

NARA crawled all federal Web pages to create a snapshot of federal Web sites in 2000, 2004, and 2006 (http://www.Webharvest.gov). Anticipating a possible change in the Congressional leadership or the administration, NARA chose major election years to capture federal Web sites. There was an expectation that another capture would be done in 2008 after the presidential elections. However, in the spring of 2008, NARA issued statements that limited its involvement in saving federal government Web sites (NARA 2008). NARA emphasized internal agency responsibility and questioned the usefulness of capturing only a one-day snapshot. However, there was enough external pressure and public concern that a private-public consortium was formed to capture all pages under the .gov domain [Library of Congress, 2008]. For the moment, it seems the government is still struggling with who is responsible for aging content and which information needs to be preserved.

Research in aging digital content naturally crosses several disciplines: library science, archival studies, information retrieval, database management, and computer networks. The existing infrastructure to support paper and book objects moves in a complete lifecycle. The problem of aging content on the Web reveals that digital objects need similar lifecycle and infrastructure support. Cyber-infrastructure is a combination of technology, users and creators involved in digital content collections. Research programs that emphasize cyber-infrastructure, like the National Science Foundation, cope with the complex interconnections of data creation, data collaboration and data use. Cyber-infrastructure research approaches are naturally holistic and may be better suited to discern how content can efficiently move through all stages.

Specific research questions related to this area include:

- If librarians and archivists guide print objects through the lifecycle, who will be the equivalent leaders in the digital realm?
- While cultural institutions clearly took the lead for printed objects, digital objects often have business, copyright and trademark restrictions and profit incentives. What types of institutions will be establishing best practices?
- Authors are not involved in the process of getting their books on or off library shelves. Will digital creators have to take a more active role in the long-term fate of their work?

Past research has emphasized the early stages of the life cycle: creation and retrieval of digital content. Future research needs to move towards the final stages in the digital lifecycle: removal and storage. Joint venture research between librarians and information systems researchers may result in better understandings of what to keep and how to preserve it.

VII. SUMMARY

The Web has grown since its inception, doubling in size every few years. Along with the growth has emerged the phenomenon of outdated content on the Web, with several estimates pointing out that large fractions of the Web are stale. Given the importance of the Web in today's decision making, the obsolescence phenomenon has to be investigated in further detail. To this end, the panel discussed the phenomenon, fusing a wide range of perspectives and articulating a set of core issues that need further research investigation.

One primary issue that arises is the definition of obsolescence. While it is tempting to define obsolescence using a metric such as age, it is more probable that the definition is multiply determined by factors such as scientific discovery and the community in question. The definition is crucial to the field of information life-cycle management. The next issue arising out of the definition is the rate of content aging, involving factors such as document dormancy, logical connectivity and demise of the content author. It is improbable that either tooling or manual editing can provide a satisfactory solution, leading to the opportunity for a more systemic solution in this area, particularly in the sensitive area of medical information management. A related issue is the relationship of content aging to Web standards and implementation technologies and the related legal implications, all of which are complicated by the emergence of dynamic content and aging issues. Next and importantly, since search engines determine document relevance using content age, the impact of content aging on information search as an input to strategic decision making needs to be explored in greater detail. Finally, the panel compared this digital problem to that in the analog world using the library as an analogy, exposing similarities and differences.

ACKNOWLEDGEMENT

REFERENCES

EDITOR'S NOTE: The following reference list contains the address of World Wide Web pages. Readers, who have the ability to access the Web directly from their computer or are reading the paper on the Web, can gain direct access to these references. Readers are warned, however, that:

1. These links existed as of the date of publication but are not guaranteed to be working thereafter.
2. The contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
3. The authors of the Web pages, not CAIS, are responsible for the accuracy of their content.
4. The author of this article, not CAIS, is responsible for the accuracy of the URL and version information.

ABC. (2002). "Dangerous Surfing Study: Drug Web Sites Provide Harmful Information," ABC News.com, http://www.cancer.med.umich.edu/news/melanoma.shtml, Jan 25, 2002 [ABC News.com, 2001].

Abels, J. E. (2008, September 8). "Inside The UAL Story Debacle," Forbes.com. http://www.forbes.com/2008/09/08/ual-tribune-bankruptcy-biz-media-cz_ja_0908ualstory2.html

Alvarez, P., I. Escalona, and A. Pulgarín. (2000). "What Is Wrong with Obsolescence?" *Journal of the American Society for Information Science* (51:9), 07, pp 812-815.

Bray, D., et al. (2006) "AMCIS 2006 Panel Report: The Web as a Digital Reflection of Reality," *Communications of the Association of Information Systems* 18(28), Nov 2006.

Brookes, B. C. (1970). "Obsolescence of Special Library Periodicals: Sampling Errors and Utility Contours," *Journal of the American Society for Information Science* (21:5), Sept-Oct, pp 320-329.

Buckland, M. K. (1992). "Redesigning Library Services: A Manifesto," Chicago: American Library Association. http://sunsite.berkeley.edu/Literature/Library/Redesigning/html.html.

Cho, J. and H. Garcia-Molina. (2000). "The Evolution of the Web and Implications for an Incremental Crawler," In Proceedings of the 26th International Conference on Very Large Databases, pp. 117–128.

Cox, R. (1994). "The Record: Is It Evolving?" The Records and Retrieval Report," *The Newsletter for Professional Information Managers* 10:3 (March 1994), p. 1–16.

DC. (2008). Accessed on September 19, 2008. See [http://dublincore.org/].

Dhamija, R., and J. D. Tygar. (2005). "The Battle Against Phishing: Dynamic Security Skins In SOUPS 2005," Proceedings of the 2005 ACM Symposium on Usable Security and Privacy, ACM International Conference Proceedings Series, ACM Press, July, pp. 77-88.

DRA. (2008). Accessed on September 10, 2008. See [http://www.dralegal.org/cases/private_business/nfb_v_target.php].

Egghe, L., and R. Rousseau. (2000). "Aging, Obsolescence, Impact, Growth, and Utilization: Definitions and Relations," *Journal of the American Society for Information Science* (51:11), 09, pp 1004-1017.

E Government Act of 2002. 44 U.S.C. § 101 (2002). http://hdl.loc.gov/loc.uscongress/legislation.107hr2458.

Eysenbach, G., J. Powell, O. Kuss, and E. R. Sa. (2002). "Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review," *Journal of the American Medical Association* (287:20), pp 2691-2700.

Floridi, D.L. (1996). "Brave.Net.World: The Internet as a Disinformation Superhighway?" *The Electronic Library* (14:5).

Fox, S. (2008). "The Engaged E-Patient Population," *Pew Internet & American Life Project*.

Google. (2008). We knew the Web was big . . . [Blog] 2008 July 25 [cited 2008 September 20]; Available from: http://googleblog.blogspot.com/2008/07/we-knew-Web-was-big.html.

Hensen, S. L. (2001). Sunday, December 16. "The President's Papers Are the People's Business," *The Washington Post*, p. B30.

Hensen, S. L. (1989). *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries* – 2nd ed. (Society of American Archivists).

Imam, I. F. (1999). "Multiple Approaches to Intelligent Systems, Chapter on WebMaster: Knowledge-Based Verification of Web-Pages," ISBN 978-3-540-66076-7.

Jacobs, J. A., James, R. Jacobs, and Y. Shinjoung. (2005). "Government Information in the Digital Age: The Once and Future Federal Depository Library Program," *The Journal of Academic Librarianship* 31 (3), 198-208. http://dx.doi.org/10.1016/j.acalib.2005.02.006.

Jakobsson, M., A. Tsow, A. Shah, E. Blevis, and Y. Lim. (2008). "What Instills Trust? A Qualitative Study of Phishing, Financial Cryptography and Data Security," *Lecture Notes in Computer Science*, Volume 4886/2008.

Joachims, T. (2007). "Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search," *ACM Transactions on Information Systems* (25:2), pp 1-27.

Jurinski, J. J. (2003). "Bankruptcy Step-by-Step," *Barron's Educational Series*.

Koehler, W. et. al. (2004). "A Longitudinal Study of Web pages Continued: A Consideration of document Persistence," *Information Research*, Vol. 9 No. 2, January 2004. [http://informationr.net/ir/9-2/paper174.html].

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Library of Congress. (2008, Aug 14). "Library Partnership Preserves End-of-Term Government Web Sites," from http://www.loc.gov/today/pr/2008/08-139.html

Line, M. B. (1993). "Changes in the Use of Literature with Time - Obsolescence Revisited," *Library Trends* (41:4), pp 665(619)-684.

Maurer, H. (1998). "Web Based Knowledge Management," *Computer*, March.

McArthur, T. (1986). *Worlds of Reference: Lexicography, Learning, and Language from the Clay Tablet to theComputer*. New York: Cambridge University Press.

"Melanoma Web Sites Give Incomplete or Wrong Information" [http://www.cancer.med.umich.edu/news/melanoma.shtml, Jan 25, 2002]

NARA. (2008, April 15). "National Archives and Records Administration Web Harvest Background Information," from http://www.archives.gov/records-mgmt/memos/nwm13-2008-brief.html

National Library of Medicine. (2007). "Developing Permanence Levels and the Archives for NLM's Permanent Web Documents," 2007 November 6 [cited 2008 August 2]; Available from: https://www.nlm.nih.gov/psd/pcm/devpermanence.html.

Nelson, A. K. (2002). "Now, More Than Ever, We Need to Know the History Of Our Recent Past," *Academe* 88 (4), 59.

Nelson, M. L., and B. D. Allen. (2002). "Object Persistence and Availability in Digital Libraries," D-Lib Magazine, Volume 8, Number 1, January.

Nicholas, D., P. Huntington, T. Dobrowolski, I. Rowlands, H. R. M. Jamali, and P. Polydoratou. (2005). "Revisiting 'Obsolescence' and Journal Article 'Decay' through Usage Data: An Analysis of Digital Journal Use by Year of Publication," *Information Processing & Management* (41:6), 12, pp 1441-1461.

Norton Article Library, http://www.symantec.com/norton/products/library/article.jsp?aid=Web_doctor, Sep, 2007.

Owyang, J. (2008). "Social Network Stats: Facebook, MySpace, Reunion," 2008 January 9 [cited 2008 September 20]; Available from: http://www.Web-strategist.com/blog/2008/01/09/social-network-stats-facebook-myspace-reunion-jan-2008/.

Rabinovich, M., and O. Spatscheck. (2002). *Web Caching and Replication*. Boston, MA: Addison-Wesley.

Rainie, L., and J. Shermak. (2005). "Search Engine Use November 2005," P*ew Internet and American Life Project* and comScore Media Metrix.

RDF. (2008). Accessed on September 20, 2008. See [http://w3.org/TR/rdf-syntax-grammar/].

Ritzenthaler, M. L. (1993). *Preserving Archives and Manuscripts*. Chicago.

Rowe, N. C. (2005). "Types of Online Deception," in: *Encyclopedia of Virtual Communities and Technologies*. Hershey, PA: Idea Group.

Samuelson, P. (1998). "Encoding the Law into Digital Libraries," *CACM: Communications of the Association for Computing Machinery* 41 (4), 13-18. http://doi.acm.org/10.1145/273035.273039.

Sandborn, P. A., F. Mauro, and R. Knox. (2007). "A Data Mining Based Approach to Electronic Part Obsolescence Forecasting," *IEEE Transactions on Components & Packaging Technologies* (30:3), 09, pp 397-401.

Section 508. (2008). Accessed on September 12, 2008. See [http://www.section508.gov/].

Shekelle, P. G., et al. (2001). "How Quickly Do Guidelines Become Outdated?" *Journal of the American Medical Association* 286(12): p. 1461-1467.

Shivakumar, N. and H. Garcia-Molina. (1998). "Finding Near-Replicas of Documents on the Web," in *World Wide Web and Databases International Workshop WebDB'98*, pp. 204-12, Valencia, Spain, March.

Simon, H.A. (1981). *The Sciences of the Artificial* (2nd. ed), Cambridge: MA, MIT Press.

Smith, W. (2008). "Conformance with Web Accessibility and Technical Recommendations: Lessons Learned from California," (Doctoral dissertation, Claremont Graduate University.

Sustersic, J. P., A. R. Hurson, and R. M. Nickel. (2006). "An Analysis of Internet Data Update Behaviors," *20th International Conference on Advanced Information Networking and Applications (AINA'06)*: IEEE.

Tallon, P. P., and R. Scannell. (2007). "Information Life Cycle Management," *Communications of the ACM* (50:11), 11, pp 65-69.

Technorati. State of the Blogosphere. (2008). 2008 [cited 2008 September 23]; Available from: http://www.technorati.com/blogging/state-of-the-blogosphere/.

The Internet Archive. http://www.archive.org/index.php.

Thibodeau, K. (2002). "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," in *The State of Digital Preservation: An International Perspective*, Washington, DC: Council on Library and Information Resources.

University of Michigan News Archive. http://www.cancer.med.umich.edu/news/melanoma.shtml, Jan 25, 2002.

van Harmelen, F. and J. van der Meer. (1999). "WebMaster: Knowledge-Based Verification of Web-Pages in Multiple Approaches to Intelligent Systems," Springer Berlin: Heidelberg, Germany. p. 256-265.

Venkatsubramanyan, S., and S. K. Kwan. (2008). "A Web Search Model for Strategic Decision Making," *AIMS International Journal of Management*, Volume 2, Number 3, September, pp. 101-118.

Wang, J., N. Xiao, and H. R. Rao. (2008). "An Exploration of Online Information Search Behavior Using Search Engine Log," Arlington, TX: College of Business Administration, The University of Texas at Arlington.

W3C. (2008). Accessed on September 19, 2008. See http://www.w3.org.

WCAG. (2008). Accessed on September 18, 2008. See http://www.w3.org/WAI/intro/wcag.php.

ABOUT THE AUTHORS

Brent Furneax is a doctoral candidate in Information Systems at York University's Schulich School of Business. His current research interests include the processes surrounding individual and organizational decision making, the strategic management of organizational knowledge, and questions related to the end of the information system life. He is currently pursuing dissertation research that seeks to better understand the factors that drive and constrain organizational decisions to discontinue their use of information systems.

Dr. Timothy R. Hill is the chair of the Management Information Systems department in the College of Business at San Jose State University. Research interests focus on innovative applications of emerging technologies and have spanned artificial intelligence, neural networks, and graphic visualization and manipulation; and more recently, mobile, multi-layered, media-rich e-learning environments, and digital reality impacts on identity and perception.

Wayne Smith is a Lecturer in the Department of Management at California State University, Northridge. He received a Ph.D. in Information Systems and Technology from Claremont Graduate University. He has worked and consulted in the management of information technology since 1984. He has worked in both the public and private sector. He has presented or published in the areas of strategic management, broadband (xDSL), diffusion of XML technologies, web accessibility, and reflective learning.

Dr. Shailaja Venkatsubramanyan is an assistant professor in the Management Information Systems department in the College of Business at San Jose State University. She has a Ph.D. in MIS from the University of Arizona and has worked at Price Waterhouse, Tulane University, and Kanisa (a Knowledge Management startup in the Silicon Valley). Her research interests include information retrieval, knowledge management, search engine valuation, Web log analysis, and impact of online information on impression formation.

Dr. Jingguo Wang is an assistant professor in the University of Texas at Arlington. He graduated from SUNY-Buffalo. His work has been published or forthcoming in IEEE Transactions on SMC, EJOR, ISR and other journals, and received best paper awards at AMCIS and the International Conference on Internet Monitoring and Protection. His research interests include information assurance, decision making, and online information search behavior.

Anne Washington is a Ph.D. candidate at the George Washington University (GWU) School of Business in the Information Systems and Technology Management doctoral program. She holds a Bachelor's of Arts (BA) in computer science from Brown University and a Master's in Library Information Science (MLIS) from Rutgers University. After working for software development firms including Claris Software division of Apple Computers, she worked for Barlcays Global Investors in multiple technology and corporate training positions. Since 2001, she has been working for the Congressional Research Service at the Library of Congress as an information systems librarian.

Paul Witman is an assistant professor of Information Technology Management in the School of Business at California Lutheran University. Witman holds a Ph.D. in Information Systems and Technology from Claremont Graduate University. His research interests include information security, usability, technology adoption and continuance, and electronic banking and finance.