The World Wide Web: from one proposal for an information management system to an ever-widening universe of information yet to be discovered

Authors: Martínez Méndez, F. J.[1]; Pastor Sánchez, J.A.[2]; López Carreño, R.[3] y Rodríguez Muñoz, J.V.[4] Departamento de Información y Documentación de la Universidad de Murcia.

Resumen

Este documento es una reflexión sobre la interrelación entre la gestión de la información en las organizaciones y los sistemas de hipertexto, un vínculo cada vez más conocido desde la propuesta de creación de la WWW realizada hace casi treinta años por Berners-Lee. Se lleva a cabo una revisión de los postulados teóricos del hipertexto, proporcionados principalmente por Bush y Nelson, como una posible solución a la crisis de la gestión de la información, comparándose con la situación de los primeros años noventa en las organizaciones. Se estudian los problemas detectados por Berners-Lee en su entorno de trabajo que le llevaron a desarrollar el WWW, la evolución que ha tenido esta tecnología y la forma en que todavía se necesita para una gestión adecuada de la información. El grado actual de desarrollo de la tecnología permite augurar una buena gestión de los contenidos de las organizaciones. Los problemas ahora están más relacionados con la actitud de los responsables de las organizaciones. Es necesario intentar que las organizaciones apuesten decisivamente por estas tecnologías para aumentar su efectividad en el momento de la gestión de la información. Además, la WWW es un sistema en continua evolución en el que debemos continuar investigando, aún no hemos llegado a vislumbrar el final de esta tecnología.

Palabras clave

Gestión de la Información, World Wide Web, Hipertexto

Abstract

This work is a reflection on the interrelation between the management of information in organizations and in hypertext systems, in the sense that it is a creation of the WWW. A review of the theoretical postulates of hypertext, proposed mainly by Bush and Nelson, is presented as a possible solution to the crisis of information management, comparing with the situation of the early nineties in organizations. We study the problems detected by Berners-Lee in his work environment that led him to develop the WWW, the evolution that this technology has had and how it is still needed for an adequate management of information. The current degree of technology development allows good management of the contents of organizations. The problems are now more related to the attitude of those responsible for the organizations. It is necessary to try that the organizations bet decisively for these technologies to improve their functioning at the moment of the management of the information. In addition, the WWW is a system in continuous evolution in which we must continue to investigate, we have not yet reached a final vision of this technology.

Keywords

Information Management, World Wide Web, Hypertext

[1] javima@um.es
[2] pastor@um.es
[3] rosanalc@um.es
[4] jovi@um.es

Introduction

The improvement of organisational information management has always been an area of great concern to our discipline and one of its core drivers from the very beginning. Had this not been the case, how would Callimachus of Cyrene ever have been able to manage the thousands of papyrus scrolls that filled the shelves of the Library of Alexandria? If no form of information management had existed at that time, he might have resorted to magic – an art the ancients may have been well versed in but has significantly fewer adepts today. Information management as a scientific discipline has been defined as 'the application of management principles to the acquisition, organization, control, dissemination and use of information relevant to the effective operation of organizations of all kinds' (Wilson, 2002). As information management is a core function of all organisations, an organisational perspective is essential to any discussion of this subject (Kirk, 1999). According to Wilson (2002), the term *information* covers any type of information of internal or external origin that may be of value to an organisation, including data resources concerning such matters as production and turnover as well as records and files related to areas such as personnel functions, market research and competitive intelligence drawn from a wide spectrum of sources. Thus understood, information management is an area that concerns to the value, quality, ownership, use and security of information in the context of organisational performance, 'information management is contextualised by the organisation' (Kirk, 1999).

Information management gained a level of visibility in the 1970s when the US National Commission on Federal Paperwork was created to explore the possibilities of reducing the burden of corporate costs related to government-required paperwork in that country (Wilson, 2002). The work of this commission, which issued as many as twenty reports on the subject, provided a framework for development of the discipline (Trauth, 1989). By the early 1980s it was clear that information management going forward would necessitate the greater implementation of computer technology and the establishment of policies related to records retention. Trauth (1989) cites a number of the pioneers such as Horton, Synnott and Gruber whose thinking would drive the convergence of approaches to information management that marked that decade. Horton, for example, asserted that information overload posed as great a problem as the loss of information. Synnott and Gruber, on the other hand, spoke of what they perceived as a natural progression from data processing to a more centralised form of information management under a chief information officer, a title that has since been superseded in many spheres by the term *information governor* (Kooper et al., 2011).

The purpose of information management is not only to preserve documents organisations generate on a daily basis for the benefit of researchers hunting for data they need to write a doctoral thesis or historical novel. The goal of today's information management professionals is to analyse information, describe it properly in line with internationally accepted standards, define its life cycle, create backup copies of whatever documents could have lasting historic value, establish a

policy as to when and in what form information should be made available to the public and, finally, delete documents once it is no longer needed. Kirk (1999) says that 'information and his management contributes to the achievement of organisational goals' and, in this line, Picot (1989, 238) says 'information has become the prime production factor, and the purpose of information management is to make sure that it is put to good use at both strategic and operational levels'. Wilson aptly pointed out (1989) such this management were not possible without an adequate technological support framework that allows for the establishment of workflow systems and levels of access to document archives.

The implementation of these ideas and technologies in the information management carried out by any type of organisation or applied to any kind of information (such as statistics on bus ridership at a specific hour, designs for tables and chairs, sales data on a certain model of automobile for a given urban area or age bracket or documentation of projects carried out in the past for a customer), which constitutes the fulfilment of the third aspect of modern information management outlined above, it is an activity that can and should be considered strategic in the business environments and situations (Andersen et al, 1994) (Winter et al., 2001) (Myburgh, 2002) (Earl, 2012). This key process requires professionals who not only have a good grasp of document management techniques and information and telecommunications technology but also a thorough understanding of the organisations they work for and the nature and purpose of the documents in their care. Such individuals must have a proactive attitude regarding their organisations' internal cultures as well as the environments in which they operate and be capable of analysing needs and information flows and identifying internal and external problems and issues related to competitors. They must also reject the ingrained notion that their role is to provide information only on request, taking a user-oriented approach and openly offering assistance to colleagues (Dervin and Nilan, 1986). Information managers must help create an atmosphere in which information and its management contribute to the achievement of their organisations' objectives (Kirk, 1999).

Hypertext as an information management solution

As the Second World War drew to a close, it became clear that the information management systems then in place did not allow for the adequate classification and organisation of the burgeoning volume of information that needed to be handled, which increasingly included large sets of documents presented in the form of graphics, sound recordings, design layouts, maps or photographs rather than text. Documents microfilmed by US armed services personnel in archives and at industrial sites in Germany accounted for much of this material. As attempts to classify and index these documents by means of traditional systems repeatedly failed, experts began to explore alternatives. The new proposed systems provided a better view of the information they contained in that they fostered associations between ideas and concepts introducing the non-sequential acquisition of knowledge. Vannevar Bush (1945) proposed the development of what he referred to

as a '*memex'*, a device he conceived as a vast database capable of storing any type of document. This machine was to have consisted of a desk fitted out with a keyboard and a set of levers by means of which a user would be able to view microfilmed data projected on a transparent platen allowing users to add their own marginal notes and comments offering the reader the option of becoming an author as well – a previous precedent of the groundbreaking dualistic character of Web 2.0 publishing (O'Reilly, 2005). This machine was never built but it fuelled the idea that new approaches to be explored and underlined the futility of continuing to rely on outmoded information management systems. Theodor Nelson coined the term *hypertext* in order to mean a body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represented on paper (Nelson, 1981; Conklin, 1981). *Xanadu*, the system he set out to create, can be considered as forerunner of the World Wide Web as we know it today. Nelson envisioned organising discrete pieces of information within a vast global network of scientific knowledge in a non-sequential manner so as to allow users located anywhere to incorporate them into documents of their own creation by means links rather than traditional copy and paste techniques. The main objective of the *Xanadu* project was to build a hypertext server accessible from any computer and capable of storing and linking the world's entire literary heritage. Nelson's ultimate goal was to create a global library of literary works stored on personal computers but simultaneously available to an unlimited number of users. These documents were to receive unique identifiers similar to the URIs and URLs assigned to every item published on the WWW. Ath the same time, this documents would be stored in much the same way that files are currently stored on computers linked up to P2P networks and accessed by means of search applications and file sharing, fulfilled the requirements of what Berners-Lee and other internet gurus refer to as the *decentralized web* (Hardy, 2016).

These ideas have been impossible to realise in the 1960s (as Bush's in the 1940s), was similar to the WWW as we know it today apart from one important exception: Nelson envisioned a hypertext system that would allow users to create documents containing links to snippets of text previously produced by other authors. Visualisation of these hybrid edited documents hinged on a *xanalogic* storage technique (Nelson, 1999) that the inventor attempted to develop at a time when computers had extremely limited memory storage capacity. It is somewhat surprising that Nelson's concept has not been implemented in the current WWW, an environment notable for its users' (often excessive) proclivity for copying and pasting fragments of existing texts. It would not be unreasonable to say that Nelson's hypertext system has yet to be fully developed.

It is easy to understand the frustration of an inventor whose ongoing work was overshadowed in the early 1990s by the ideas of a young British researcher who worked in the European Organization for Nuclear Research (CERN). When web technology began to gain traction, a number of writers seeking to acknowledge Nelson's contribution to its development billed it 'WWW, Xanadu at last', an homage that seems not to have fully satisfied the inventor, who has publically

recognised the success and importance of the Web but described it as a 'brilliant simplification' of the full hypertext environment that he had envisioned (Nelson, 1999, 1), (Whitehead, 1996), (Nelson, 2007). The failure of the technology implemented over the past fifty years to facilitate effective information management systems in the organisations has been especially doleful for researchers who need to draw upon a wide range of sources and keep track of a number of simultaneous projects. It struck Berners-Lee as incredible that a 'marvellous organisation' like CERN, which brings together so many people of high intellectual level (many of them very creative) could be plagued by such a problem.

CERN's researchers were nominally organised within a hierarchical management structure that did not restrict the ways in which they communicated or shared information, equipment and software. In practice, the organisation's working structure resembled a web with multiple interconnections that evolved over time rather than a conventional hierarchical system. People new to the institution as well as those assuming new responsibilities were given basic advice about whom they might contact regarding available facilities and subsequently left to figure out the organisational aspects of their work more or less on their own. The tremendous success of CERN's projects invariably offset any organisational shortcomings (Berners-Lee, 1989). Duplication of information acquisition efforts caused by lapses in internal knowledge sharing and the potential loss of information gathered were not considered to be particularly important; technical details of prior projects were occasionally lost forever and files gone missing were often hurriedly tracked down by means of intuitive detective methods. CERN inevitably considered these problems as the more imponderable facets of its ongoing endeavours. Berners-Lee (1989) has cited the centre's high researcher turnover rate as a factor contributing to the risk of information loss. As most residencies last approximately two years, preventing the loss of information attributable to churn is an ongoing challenge. Making sure that new personnel know the ropes (especially at the most advanced levels of research activity) is an ongoing task that consumes much of his and others' time. Whenever the pace of new arrivals rises, control over information management slips proportionally.

At a certain point, Berners-Lee identified another problem that had previously gone undetected: the way in which documentation related to a project was being registered. If a CERN experiment were a static, one-off event, all related information could be documented in a single book for later reference. However, this was the exception rather than the rule at CERN, where the focus of organisational and research interest constantly changes in response to the emergence of new ideas and technologies and the need to overcome unexpected technical glitches. Researchers needing to make small project adjustments that theoretically would affect only one small area within the organisation were expected to verify and report whether such changes might have an impact on other areas and projects. Maintaining a centralised book up to date was a time-consuming and impractical endeavour. As researchers found it difficult to obtain information regarding issues such

as where a module was used, who had written a given code or where that particular person worked, they tended to rely on verbal clarifications provided by colleagues. Conventional information systems employed at the time were particularly ill suited for managing software and related documentation: 'it is a type of application that is not well supported by traditional information management systems' (Jarwa and Bruandet, 1990, 69).

Owing to the limitations of CERN's document management system, workers at the centre tended to pass only tacit knowledge (often in fragmentary form) along to colleagues. Little of the information shared tacitly between co-workers was ever formally articulated. This practice dangerously limited the generation and flow of the kind explicit knowledge a high-level research organisation such as CERN needs to maintain to preserve its competitive edge (Nonaka, 1994, 16-17). The situation grew worse whenever researchers worked at the centre for only short periods of time, and the rapid turnover of personnel threatened to choke off the flow of the organisation's internal knowledge transmission, the organisation needed to resolve the problem of information losses rather than write them off as isolated incidents. Although Berners-Lee (1989) believed that commercial solutions would be available within a decade, he realised that the current situation was becoming untenable, he wanted a tool that would facilitate the CERN's work going forward: a new information management system that would grow and evolve in step with the organisation and its projects, considering that the survival of any research organisation is irrevocably linked to its ability to improve its information management (Earl, 1994) (Huotari, 1995) (Burks, 2006) (Middleton, 2007).

To be up to this task, the storage method implemented must allow for any type of information to be entered. For an organisation like CERN, a *web* of notes with links that served as references was much more useful than a fixed hierarchical system of files organised in nested folders by an operating system's file manager. Many people tend to describe complex systems by means of diagrams containing circles and arrows, which are more useful for mapping out the relations between things than tables and directories. If we think of the arrows as links and the circles as nodes (and further conceptualise these nodes as short notes or pieces of information such as articles, summaries and comments), it is easy to mentally construct a constantly evolving linked system that connects people with bits of information. Once such a scheme is implemented, project information is longer stowed away in a single folder of little use to a newly arrived researcher but rather forms a part of an organisation-wide information network that fosters links between a much greater spectrum of people and departments. This approach to storage clearly favours the survival of information and firmly establishes the idea of hypertext as an information documentation system. According to Pastor-Sanchez and Saorin-Perez (1995), 'something more than a simple tool for organising information for subsequent retrieval, hypertext should be considered an active structure capable of satisfying information needs by means of the creation of hyperdocuments in which related information is condensed within a hypertext network'. In any case, almost all of the

users of the CERN information system are scientists skilled at retrieving information from libraries, and as Conklin has noted (1987, 37), accustomed to navigating through it in ways that mirror the characteristics and functions of hypertext:

> Intertextual references are not new. The importance of hypertext is simply that references are machine-supported. Like hypertext, traditional literature is richly interlinked and is hierarchically organized. In traditional literature, the medium of print for the most part restricts the flow of reading IO follow the flow of linearly arranged passages. However, the process of following side links is fundamental even in the medium of print. In fact, library and information science consist principally of the investigation of side links. Anyone who has done research knows that a considerable portion of that effort lies in obtaining referenced works, looking up cross-references, looking up cross in a dictionary or glossary, checking tables and figures, and making notes on notecards. Even in simple reading one is constantly negotiating references to other chapters or sections (via the table of contents or references embedded in the test), index entries, footnotes, bibliographic references, sidebars, figures, and tables. Often a text invites the reader to skip a section if he is not interested in greater technical detail.

The Web solution

During this period (in the last 80's), scientists relied heavily on UUCP network architecture to disseminate information categorised into *newsgroups* (forerunners of today's online discussion groups). However, initial discussion topics inevitably spawned a succession of parallel, interrelated topics that the hierarchical tree structure underpinning UUCP protocols was incapable of managing (another weakness of newsgroup architecture that made it hard to use was its lack of a search mechanism). Some way of making it easier to retrieve and read posted texts needed to be developed, several potentially suitable hypertext systems had already been developed (Conklin, 1987, 27-32), the most notable being Guide, HyperCard and Xerox's Notecards. During his first residency at CERN in the early 1980s, Berners-Lee had written the code for Enquire, a program similar in many ways to the commercial systems mentioned earlier that 'allowed to fill a document with words that, when clicked, would lead to other documents for elaboration' (Wright, 1997), in order to record the progress of software projects.Once Berners Lee returned to CERN almost a decade later, everything at the organisation had changed: the aforementioned software applications (particularly HyperCard) had been broadly implemented and researchers were using Internet networks in a very different way.

> The world had changed. The Internet, though still unknown to the public, was now firmly rooted. It was essentially a bare-bones infrastructure, a trellis of empty pipes. There were ways to retrieve data, but no really easy ways, and certainly nothing with the intuitive, neural structure of hypertext. (Wright, 1997)

The time had come to tackle the centre's information management problems, and the answer would come in the form of a Web-based hypertext information system. Berners-Lee was not interested in halfway measures based exclusively on previous knowledge and systems; he was looking for a solution that took advantage of prior achievements but also exploited the possibilities of newer ideas and technology. The first step was to create a simple protocol for navigating the Web (http) and a mark-up language suitable for web communication (HTML). He subsequently focused on solutions for support (a web server), navigation (web client software) and the edition of texts using HTML. The outcome of this endeavour was a revolutionary information system that has changed the world as we know it and quite possibly constitutes the greatest technological innovation ever to have been implemented in our field. Its importance to information management nevertheless pales in comparison to the changes its introduction has wrought over the past few decades. The Internet has driven the creation of a virtual sphere of practically limitless information accessible to people all over the world regardless of their geographic location and the limitations of local infrastructures and technology.

One may wonder if Berners-Lee was initially aware the potential scope and ramifications of his creation (it is a question well worth asking). The impact of the Web and web-related technologies on the daily lives of a vast number of people around the globe today is however an irrefutable fact. Those young enough to be digital natives cannot imagine life without the World Wide Web, especially since the massive popularisation of applications such as email, search engines, messaging systems and the interchange of photographs, opinions and comments via social networks. Acceptation of the Web was swift and definitive. Within months of its initial introduction, it eclipsed Gopher, an information system protocol then popular with university communities that functioned like a message board, permitted connection between networked servers and provided gateways to OPACS modules.

The user experience and wide spectrum of features offered by the Web made it a memorable object of love at first sight for those who had an opportunity to try it out. Leonard Kleinrock, one of the inventors of the Internet, commented in an interview that what he enjoyed most about his involvement in the project was 'having been there' while it all was happening. Before Internet access became common, many professors had to explain how information was retrieved from online databases such as Medline, ERIC and Lisa by means of slide presentations or demo software owing to the prohibitively high costs of dial-up connection to these service providers. Improved Internet connections on university campuses and the emergence of the Web brought professors and students closer to the formerly inaccessible information industry, making it possible for them to make modest contributions to development in the field and giving them a deeper understanding of the technology that underpinned it.

The Web was created at a time when Internet access was improving substantially in many parts of

the world. This coincidence has given rise to the widespread (but not completely erroneous) notion that the Web was responsible for introducing the Internet into our lives. It was an exceptional period during which information professionals worked hand-in-hand with professors and researchers to map out the potential uses and applications of the Web, actively participating in its development of the technology rather than leaving the task exclusively in the hands of programmers and technicians. They collaborated on the construction of early web pages (the design of which left much room for improvement) and began to give form to the first websites – efforts that over time led to the integration of information, services and applications on web portals. This process of consolidation took several years; the time required was different for every country and type of organisation. Communication via the Web during this initial phase was unidirectional. Author/editors provided information for user/readers. The incorporation of user feedback and interactivity was yet to come. These primitive websites, now classified as Web 1.0 (O'Reilly, 2005, 226) (Cormode & Krishnamurthy, 2008), usually consisted of a single page containing links to other pages the author believed to be of interest to readers. This approach to information management, which represented the adaptation of a reference service traditionally provided by libraries, constituted a springboard for the development of the first information retrieval systems to appear on the Web, which were indexes and directories such as Yahoo! Directory (secondary sources of information – indexes or classified lists – that route users to primary sources) a task that information retrieval systems have always performed and continue to perform today. Although they hark back to Berners-Lee's original idea of improving researchers' access to resources, as Mutula and Wakumoya (2007, 152) have noted, online web-based reference sites offer more than reference services.

> Moreover, web-based reference services can provide support for users who find online tools and resources unfamiliar, difficult to learn, or insufficient to answer their information needs. In addition, web-based reference services can provide valuable user feedback to collection builders, which may help them to tailor their resources better and maximise their investment in content creation. Through the intervention of the technology, more so the Internet and the Web, information seekers can access information resources from outside of the physical library.

The dovetailing of reference services management and web technology gave rise to the concept of digital libraries: 'the logical extensions and augmentations of physical libraries in the electronic information society. Extensions amplify existing resources and services and augmentations enable new kinds of human problem solving and expression' (Marchionini, 1998).

As content production and publication via the Web increased, the manual management of these information retrieval systems became so unwieldy that material published could only be reviewed in a cursory fashion. This dilemma led to the development of search engines, systems that feed their databases by means of *crawler* programs that methodically scan their way through Web

pages, copying and automatically indexing the hyperlinks and content they contain (Schwartz, 1998, 974-975). These systems conduct a more thorough review of material published on the Web than would be possible using manual methods and constantly capture updates of web page content and documents they have previously scanned. Search engines represented a great leap forward when they were introduced because they obviated the need to acquire advanced information retrieval language skills. For the very first time, Internet users were able to access large volumes of documents simply buy entering a few key search terms into a search box. Although users reacted enthusiastically to the introduction of search engine technology, these systems also had their critics. As the algorithms they used had been developed prior to the creation of the Web, there were issues regarding the process of refreshing indexes that had yet to be worked out. They were also incapable of indexing an entire website. Given these shortcomings and other problems intrinsically linked to the early Web environment, it is not surprising that many had serious doubts as to the effectiveness of information retrieval performed by these systems. As Landoni and Bell (2000) have noted:

> The Web is a huge, open, multicultural, multilingual, almost uncontrolled and ever growing repository of information. One of the most serious problems faced by web users at the moment is to be able to retrieve valuable information by fishing it out of a huge sea of neither regulated, not guaranteed, dynamic data.

If users at the time had misgivings about the effectiveness of search engines, they also had their doubts about the traditional evaluation methods they employed, the majority of which relied on relevance judgements (Landoni and Bell, 2000, 124). These reservations provided the impulse for a wide range of studies that proposed better methods of evaluating search engine effectiveness (Oppenheim, et al., 2000), (Martínez Méndez, 2001), (Martínez Méndez and Rodríguez Muñoz, 2003). Nevertheless, mistrust concerning how these systems functioned was quickly dispelled when two whiz kids at Stanford University created Google, a search engine that all immediately after its launch proved its superiority in terms of document indexing and information retrieval capacity. What underpinned Google's extraordinary performance? Like other individuals throughout the history of Web development, its creators drew inspiration from the information sciences. PageRank, Google's result alignment algorithm (Page et al., 1999), is an updated version of the impact factor, a system developed by Gardfield (2006) based on the computation of article citations that has been used for many years to rank scientific publications.

> There has been a great deal of work on academic citation analysis that has published an interesting theory of how information flow in a scientific community is an epidemic process. There has been a fair amount of recent activity on how to exploit the link structure of large hypertext systems such as the web . . . It is obvious to try to apply standard citation analysis techniques to the web's hypertextual citation structure. One can simply think of every link as being like an academic citation. So, a major page like

> http://www.yahoo.com/ will have tens of thousands of backlinks (or citations) pointing to it. This fact, that the Yahoo home page has so many backlinks generally imply that it is quite important. Indeed, many of the web search engines have used backlink count as a way to try to bias their databases in favor of higher quality or more important pages. However, simple backlink counts have a number of problems on the web. Some of these problems have to do with characteristics of the web which are not present in normal academic citation databases. (Page et al, 1998, 2)

Google's new criterion for ordering search results, which took into consideration the number of links on a web page as well as their quality (a value pegged to the number of incoming and outgoing links associated with it), was an immediate hit with the majority of search engine users. However, this formula has not been the sole driver of Google's success. Other reasons for its ongoing popularity are its enhanced indexing process and indexing capacity (initially limited to a small set of characters but quickly ramped up to cover the entire content of a given page), practice of grouping results related to a particular website and automatic search suggestion feature. These aspects of the system, as well as the company's ongoing efforts to refine its search processes and the launch of additional related services, have made Google a leading (and even perhaps the world's foremost) information technology enterprise. A system designed specifically to track research projects quickly became a springboard for further theoretical research on information retrieval and new points of focus in the development of information retrieval systems. The convergence of information management and information retrieval – two fields previously considered to be quite distinct – produced some very interesting consequences, none of which are surprising in light of the fact that a document must be stored in some fashion before it can be retrieved. Storage is part of the life cycle of information, which is central to the process of information management (Wilson, 2002). Therefore, information retrieval has, in one way or another, "always been there", even though other aspects of information management may have traditionally received more attention.

Web 2.0: further opportunities for confluence

Many improvements have made to search systems over the last twenty-five years. Although users may never be conscious of the continual substitution and addition of search, indexing and alignment algorithms, they do notice the introduction of more obvious innovations such as the generation of automatic and personalised search suggestions based on their previous search queries. The environment in which search engines function today is radically different than that of 1995 and the volume of content generated has grown exponentially. The Web has become a two-way street (Bacallao Pino, 2010) in the sense that with very little effort on their part, passive users (readers) can become content editors (authors). As noted in earlier sections of this paper, this

dual capacity underpins the paradigm of what is referred to as Web 2.0 (O'Reilly, 2005), which emerged with the popularisation of blogging. A number of scientists and writers interested in disseminating their ideas and findings through channels free of the rigid constraints of traditional scientific journals began to blog, a concept originally developed to provide documental support to technology (typically software) development units. Some of the blogs they have created have become essential resources for scholars in given fields, one of which is information retrieval. Every web user is a potential author. The proliferation and unprecedented success of blogs inspired the creators of a number of web-based services we take for granted today. One was Wikipedia founder *Jimbo* Wales, who saw the potential of the culture of participation so evident in blogging communities and proved that reference sources – so important to our discipline – constitute an integral part of the Web's DNA. Another is Facebook creator Mark Zuckerberg, who perceived a widespread interest in sharing personal information easily and quickly and designed a platform to satisfy that demand.

In the light of the emergence social networking applications and the scholarly attention rightly focused on them that all outstanding issues have been duly resolved and information management and retrieval are now of secondary concern. Nothing could be further from the truth. The increasing volume of information being circulated via the Web has left us in much the same situation that Library of Congress technicians found themselves in when they took on the challenge of managing millions of microfilmed documents generated at the end of the Second World War. Today's search engines must discern between the wheat and the chaff. Among other things, they must be able to identify the original source of texts that have been copied and republished word-for-word on a plethora of other blogs and websites. A number of specialised working groups have been formed within the umbrella of the Text Retrieval Conference to address this issue:

> Our approach to the opinion retrieval task identified three aspects involved in locating opinionated blog posts: topical relevance, opinion expression, and post quality. Topical relevance is the degree to which a post deals with the given topic; this is similar to relevance as defined for ad-hoc retrieval tasks, such as many of the traditional TREC tasks. The second aspect, opinion expression, involves identifying whether a post contains an opinion: the degree to which it contains subjective information about a topic. Finally, the post quality is an estimation of the (query-independent) quality of a blog post, under the assumption that higher-quality posts are more likely to contain meaningful opinions and are preferred by users. In this last category of quality, we also include detection of spam in blogs, defining a spam blog post as a low-quality one (Mishne, 2007)

Another problem closely related to this issue is effectively conveying the purpose of a given search, something very difficult to do in two or three words. For example, a search conducted by a secondary school student looking for information s/he needs to prepare an essay on models of mixed-market economies is likely to produce (in addition to a favourably placed Wikipedia entry

whose use has been proscribed by the teacher giving the assignment) a long list of documents and sites related to university economics programmes that are of absolutely no utility to the asker. Anyone who ever been in the same situation knows what it is like to try to wade through a sea of undifferentiated and irrelevant information. Having not found a useful document quickly, a searcher is obliged to do what has become almost unthinkable: scan through several more pages of the results generated or give up and start another search.

> Searching to learn is increasingly viable as more primary materials go online. Learning searches involve multiple iterations and return sets of objects that require cognitive processing and interpretation. These objects may be instantiated in various media (graphs, or maps, texts, videos) and often require the information seeker to spend time scanning/viewing, comparing, and making qualitative judgments. Note that "learning" here is used in its general sense of developing new knowledge and thus includes self-directed life-long learning and professional learning as well as the usual directed learning in schools. Using terminology from Bloom's taxonomy of educational objectives, searches that support learning aim to achieve: knowledge acquisition, comprehension of concepts or skills, interpretation of ideas, and comparisons or aggregations of data and concepts (Marchionini, 2006, 43)

A person looking for a special type of material needs to be able to manage search results. This can currently be done by entering a search equation that prevents undesired types of documents from being included in the results. Search systems that allow a greater degree of fine-tuning allow users to narrow their searches more effectively. This is possible to do with varying degrees of success using a search tool such as Google Academic. The management of search results is a juncture of information management and information retrieval that has yet to be fully explored and developed.

The Semantic Web (Berners-Lee and Hendler, 2001) offers a new opportunity for synergy between information managers and information retrieval system developers. The initial idea of adding a semantic layer of metadata to web pages has not been a success in generalist open web environments (Zhang and Dimitroff, 2004, 311), (Phelps, 2012, 328). Only a limited number of highly specific information systems (such as those utilised by libraries and digital collections) presently use metadata schemes to provide supplementary information about document content to search engines and enhance their information retrieval capacities. This is nevertheless a good idea that is gradually catching on through the use of microformats and, more particularly, microdata (Google, 2015). Microformats and microdata have generated a great deal of interest because they make the display of search results more dynamic and appealing to users. Their potential is such that sector players have founded Schema.org, an alliance specifically devoted to the development of microdata schemes (García Marco, 2013), (Sulé, 2015). Marking up web page content with microdata is a good way of guaranteeing the prominent placement of a web page in

search results and promoting a page or site without the services of an expensive digital marketing company. It is also an information management task: further proof that the partnership between information retrieval and information management continues to be healthy and, at least in terms of this particular point, has a bright future ahead.

The nearly simultaneous emergence of the field of Linked Open Data (Bizer et al., 2008) and push for greater transparency and the public disclosure of government documents (Concha and Naser, 2012) have created another point of common interest between the two disciplines as the structure and purpose of any sets of data to be linked need to be previously defined. Much of the data in this category (such as the price of fuel commodities, which varies on a daily basis) must also be frequently updated. Linking unmanaged data is an exercise in futility.

> Linked Open Government Data represents a new data integration paradigm for sustainable growth of OGD and consequently can be considered a new Enterprise integration application approach. First, it opens up the scope of data integration from traditionally closed enterprise environments such as data warehouses to the entire Web. Users can mash up government data with crowdsourced data, privately owned data, and many other types of nongovernmental data. Second, it enables a data-oriented architecture (DOA) that decouples complex data objects into reusable fine-grained linked data on the Web. (Ding et al., 2009, 13).

The technological complexity of the Web becomes a determining factor in creating experiences for users and content and service managers alike. User experience platforms integrate a spectrum of technologies that together provide a high level of interaction between users, applications, processes, contents and, of course, other users. A UXP offers and integrated experience to both users and the managers of an organisation's digital presence on the Internet (Phifer, 2011). UXPs are useful in contexts in which information related to products, services and digital objects as well as content must be retrieved via a wide variety of platforms, browsers and devices.

We are now facing a new challenge the creators of Web 2.0 sought to address: as the number of potential search objects increases, so does the need for enhanced information retrieval capacity. Structural data plays a fundamental role in this new phase of synergy between information management and information retrieval. Search engines already incorporate techniques that improve information retrieval by means of analyses of the descriptive metadata related to web content. The idea now is to extend this focus to cover the entire body of content. The inclusion of structural data facilitates the identification of people, locations, products, prices, reviews, dates, events and much more. Metadata is currently being used to create what is referred to as rich snippets as well as in research initiatives launched by organisations such as Schema.org to improve search engine performance.

Semantic content markup is a two-way process in that datasets created from structural data drawn from content can also be used to render the markup process more effective. Web-related

information management therefore calls for the inclusion of structural data in any and all content (Pastor-Sánchez, Orduña-Malea, Saorín, 2013). This is the main focus of research being implemented in web-related information retrieval today. As a consequence, organisations must now manage their web-related information from an additional perspective. Most users manage to get by with the superficial perception that web pages are but simple containers of textual and multimedia information. However, those charged with maintaining them must understand them in terms of structural content and, at an even deeper level, as structural content that can be marked up semantically and subsequently disaggregated into component parts.

The Semantic Web provides the foundation for all of the applications described above. Open technologies and standards such as RDF, OWL, SPARQL and SKOS have been developed to enhance the interoperability of data and facilitate the direct exploitation of data sources as well as their incorporation into and extraction from web content. Information retrieval systems have become progressively more adept at generating results that respond to highly specific queries. However, data cannot be reutilised by search engines not employing proprietary software/applications and solutions such as those developed by Google and Microsoft unless it has been indexed using open and interoperable technologies. Search engines may distinguish themselves from the competition by means of the services and personalisation features they offer but not in terms of the technologies they use to position themselves. This means that organisations must align their information management processes with the fact that the integration of user experience and information retrieval now depends upon the inclusion and reutilisation of structural data by means of Semantic Web technologies and semantic markup.

Conclusions

There are many pending issues yet to be resolved concerning information retrieval on the Web. Although an IT solution may sometimes manage to mitigate (but never fully mask) the effects of inadequate previous information management, in the majority of circumstances, the success of any solution implemented will hinge to a great degree on the optimal prior treatment of the information in question. The record shows that the Web and information management have been interdependent and complementary disciplines since the time of Berners-Lee's invention of the Web and that this symbiotic relationship will remain intact for the near-term future. The introduction of Web 2.0 is changing the ways in which organisations are using the World Wide Web. Users are now playing a role in content creation and new processes, platforms and services that facilitate and rely on the reuse of content and digital objects have been developed. The retrieval of information on the Web going forward supposes meeting the challenge of developing mechanisms the effectively identify a diverse variety of content presented in the form of news, discussion forums, videos, computer files, blogs, Wiki articles, etc. This is an arduous task given the heterogeneity of Web content in terms of origin, essence, structure and level of detail.

What are the factors driving this new scenario? First of all, the introduction of technologically advanced content management systems has meant that average users now have the option to become publishers. Web content is now treated as a unit of information rather than an HTML document. Website maintenance today involves the implementation of many of the core principles of information management. Concepts and practices such as structuring, organisation, classification, the assignment of descriptive, administrative and preservative metadata, positioning, review processes and dissemination – all of which were originally developed within the field of information management – are now an integral part of web content management systems.

In a post he wrote for Google's official blog to mark the twenty-fifth anniversary of his original proposal for an internal information management proposal for CERN, Berners-Lee (2014) stated that he set out to improve the flow of information within that organisation by creating 'a web of notes with links'. He also recalled that his superiors at CERN (a centre focused heavily on physics research) considered the development of the information management system he proposed to be a low priority and believed that the organisation could not afford to devote extensive resources to a project of a generalist nature not directly related to its core mission. This decision had important consequences. Instead of a select few researchers at CERN, thousands of people throughout the world would work together to build what would become the World Wide Web, sharing information, generating billions of euros of economic value over time, transforming traditional spheres such as education and healthcare, spawning social movements and creating new habits and needs almost impossible to imagine at the outset of the project. According to the man who invented the Web, our work has only just begun. For Berners-Lee (2014), the twenty-fifth anniversary of the birth of the Web was not only a landmark to celebrate but also an opportunity to reflect, talk and act. The time to make key decisions concerning the governability and future of the Internet is fast approaching and it is of vital importance to speak of the future of the Web and attempt to answer the following questions: What can we do to ensure that the 60% of the world's population that currently does not enjoy the benefits of Internet has access in the future? What needs to be done to make the World Wide Web compatible with all the world's languages and cultures rather than only a dominant few? And last but not least, are we going to stand back and let others package and limit our online experience, which is the purpose of certain governments' and telecommunications operators' attempts to end net neutrality (Cullell March, 2012), or will we protect the magic of the open Web and the power it gives us to speak, discover and create freely?

References

Andersen, D. F., Belardo, S., & Dawes, S. S. (1994). Strategic information management: Conceptual frameworks for the public sector. *Public Productivity & Management Review*, 335-353.

Bacallao Pino, L.M. (2010) Representaciones mediáticas de las redes sociales: un estudio de casos. *Revista Latina de Comunicación Social*, 65, 114-125. Retrieved 21 June, 2015 from http://www.revistalatinacs.org/10/art/887_UZaragoza/09_Lazaro_Bacallao.html (Archived by WebCite® at http://www.webcitation.org/6m2mpm5Pj)

Berners-Lee, T. (1989) *Information Management: A Proposa*l. Retrieved 18 November, 2014 from http://www.w3.org/History/1989/proposal.html [Archived by WebCite® at http://webcitation.org/OMl]

Berners-Lee, T. and Hendler, J. (2001) Publishing on the semantic web. *Nature* 410, 1023-1024 (26 April 2001) doi:10.1038/35074206

Berners-Lee, T. (2014, 11th March) *On the 25th anniversary of the web, let's keep it free and open*. [Blog] Retrieved 18 November, 2014 from https://googleblog.blogspot.com.es/2014/03/on-25th-anniversary-of-web-lets-keep-it.html [Archived by WebCite® at http://www.webcitation.org/6m2nwuRQ0]

Bush, V. (1945, July). As we may think. *Atlantic Monthly*, 176, 101-108. Retrieved 18 November, 2014 from http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ [Archived by WebCite® at http://www.webcitation.org/6m2o9v9jU]

Bizer, C., Heath, T., Idehen, K. & Berners-Lee, T. (2008). Linked data on the web (LDOW2008). // *Proceedings of the 17th international conference on World Wide Web,* 1265-1266.

Burks, T. (2006). Use of information technology research organizations as innovation support and decision making tools. In *Proceedings of the 2006 southern association for information systems conference*, 8-14.

Conklin, J. (1987) Hypertext: an introduction and survey. *Computer*, 9 september 1987, vol. 20. p.17-41 doi:10.1109/MC.1987.1663693

Concha, G. y Naser, A. (2012) *El desafío hacia el gobierno abierto en la hora de la igualdad*. CEPAL, Naciones Unidas. Retrieved 21 June, 2015 from http://iis7-e2.cepal.org/ddpe/publicaciones/xml/9/46119/W465.pdf [Archived by WebCite® at http://www.webcitation.org/6m2ok2aSo]

Cormode, G., & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *First Monday*, 13(6). Retrieved from http://www.ojphi.org/ojs/index.php/fm/article/view/2125/1972

Cullell March, C. (2012). El futuro de la Web ante la neutralidad de la Red: estado de la cuestión en la Unión Europea. *El Profesional de la Información* 21 (1), 77-82. Retrieved from http://dx.doi.org/10.3145/epi.2012.ene.10

Dervin, B. & Nilan, M.(1986) Information needs and uses. *Annual Review of Science and Technology* 21, 3-32

Earl, M., & Khan, B. (1994). How new is business process redesign? *European Management Journal*, 12(1), 20-30.

Earl, M. J. (2012). Experiences in strategic information systems planning. Strategic Information Management, 181.

García-Marco, F.J. Schema.org: la catalogación revisitada. *Anuario ThinkEPI*, 2013, 7, 169-172. Retrieved 21 october, 2015 from http://www.thinkepi.net/schema-org-catalogacion-revisitada [Archived by WebCite® at http://www.webcitation.org/6m3mBoIGN]

Gardfield, E. (2006). The Agony and the Ecstasy - The history and meaning of the journal impact factor. *Jama*, *295* (1), 90-93. Retrieved 21 June, 2015 from http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf?wa=IPEMBI14 [Archived by WebCite® at http://www.webcitation.org/6m3mNBSXR]

Google. (2015) *Promote Your Content with Structured Data Markup*. Retrieved 21 October, 2015 from https://developers.google.com/structured-data/ [Archived by WebCite® at http://www.webcitation.org/6m3mXkCyb]

Hardy, Q. (2016) The Web's Creator Looks to Reinvent It. *The New York Times*, June 7, 2016. Retrieved 16 June, 2016 from http://www.nytimes.com/2016/06/08/technology/the-webs-creator-looks-to-reinvent-it.html?_r=0 [Archived by WebCite® at http://www.webcitation.org/6m3mj1FlB]

Jarwa, S. and Bruandet, M.F. (1990) A Hypertext Database Model for Information Management in Software Engineering. *Proceedings of the International Conference in Vienna, Austria, 1990*, 69-75 doi:10.1007/978-3-7091-7553-8_11

Kirk, J. (1999) Information in organisations: directions for information management". *Information Research*, 4 (3). Retrieved 21 June, 2015 from http://informationr.net/ir/4-3/paper57.html [Archived by WebCite® at http://webcitation.org/5v924fSQr]

Kooper, M.N., Maes, R. & Roos Lindgreen, E.E.O. (2011) On the governance of information: introducing a new concept of governance to support the management of information. *International Journal of Information Management*, vol. 31 (3). 195–200 doi:10.1016/j.ijinfomgt.2010.05.009

Huotari, M. L. (1995). Strategic information management: A pilot study in a Finnish pharmaceutical company. *International Journal of Information Management*, 15(4), 295-302.

Landoni, M. and Bell, S. Information retrieval techniques for evaluating search engines: a critical overview. (2000) *Aslib Proceedings*, vol. 52 (3), 124-129.

Marchionini, G. (1998) Research and development in digital libraries. *Encyclopedia of Library and Information Science*, vol. 63, 259-279.

Marchionini, G. (2006) Exploratory search: from finding to understanding. *Communication of the ACM*, vol. 49 (4), 41-45.

Martinez-Mendez, F.J. (2001) Aproximación general a la evaluación de la recuperación de información mediante motores de búsqueda en Internet. *Scire: representación y organización del conocimiento*. 7 (1), 11-31 Retrieved 20 November, 2015 from http://ibersid.eu/ojs/index.php/scire/article/viewFile/1139/1121 [Archived by WebCite® at

http://www.webcitation.org/6m3nVxfow]

Martínez Méndez, F.J. y Rodríguez Muñoz, J.V. (2003) Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la Web. *Information Research: an international electronic journal*, 8 (2), paper 148. Retrieved 20 November, 2015 from http://www.informationr.net/ir/8-2/paper148.html [Archived by WebCite® at http://www.webcitation.org/6m3vf0yZ4]

Middleton, M. (2007). A framework for information management: Using case studies to test application. *International Journal of Information Management*, *27*(1), 9-21.

Mishne, G. (2007) Using Blog Properties to Improve Retrieval. I*nternational AAAI Conference on Web and Social Media*. Retrieved 23 March, 2016 from http://icwsm.org/papers/3--Mishne.pdf [Archived by WebCite® at http://www.webcitation.org/6m3w42Ckt]

Myburgh, S. (2002). Strategic information management: Understanding a new reality. *Information Management Journal*, 36(1), 36-43. Retrieved from https://search.proquest.com/docview/227698180?accountid=17225

Mutula, S.M. and Wamukova, J.M. (2007) *Web information management: a cross interdisciplinary textbook*. Oxford: Chandos Publishing.

Nelson, T.H. (1981) *Literary Machines: the report on, and of, Project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom*. Sausalito, CA: Mindful Press.

Nelson, T.H. (1999) Xanalogical Structure, Needed Now More than Ever: Parallel Documents, Deep Links to Content, Deep Versioning and Deep Re-Use. *ACM Comput. Surv.* 31, 4es, Article 33 Retrieved 6 Sept, 2017 from http://www.xanadu.com.au/ted/XUsurvey/xuDation.html

Nelson, T.H., Adamson Smith, R. and Mallicoat, M. (2007). Back to the future: hypertext the way it used to be. In *Proceedings of the eighteenth conference on Hypertext and hypermedia* (HT '07). ACM, New York, NY, USA, 227-228. DOI=http://dx.doi.org/10.1145/1286240.1286303

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization science*, 5(1), 14-37.

O'Reilly, T. (2005) *What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. Retrieved 18 November, 2015 from http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html [Archived by WebCite® at http://webcitation.org/6UKWtB1Y0]

Page, L.; Brin, S.; Motwani, R. and Winograd, T. (1998) *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab. Retrieved 21 June, 2015 from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf [Archived by WebCite® at http://www.webcitation.org/6m44ff0Zk]

Pastor Sánchez, J.A. y Saorín Pérez, T. (1995) El hipertexto documental como solución a la crisis

conceptual del hipertexto. El reto de los documentos cooperativos en redes. *Cuadernos de Documentación Multimedia*, número 4, junio de 1995. Retrieved 18 November, 2014 from http://pendientedemigracion.ucm.es/info/multidoc/multidoc/revista/cuadern4/hiperdoc.htm [Archived by WebCite® at http://www.webcitation.org/6m44qavWD]

Oppenheim, C., Morris, A., McKnight, C. and Lowley, S. (2000) The evaluation of WWW search engines. *Journal of Documentation*, Vol. 56 Iss: 2, 190-211. http://dx.doi.org/10.1108/00220410010803810

Pastor-Sánchez, J.A., Orduña-Malea, E. y Saorín Pérez, T. (2013) Marcado semántico automático en gestores de contenidos: integración y cuantificación. *El profesional de la Información*, septiembre-octubre, vol. 22, n. 5, 381-391. Retrieved 16 November, 2016 from http://eprints.rclis.org/20371/1/2013%20EPI%20Pastor-Ordu%C3%B1a-Saorin%20CMS%20Semanticos.pdf [Archived by WebCite® at http://www.webcitation.org/6m454SqMS]

Phelps, T.E. (2012) An Evaluation of Metadata and Dublin Core Use in Web-Based Resources. *Libri* 62, 4, 326-335.

Phifer, Gene (2011). *Hype cycle for web and user interaction technologies*. Stamford, CT: Gartner.

Picot,

Schema.org (s.f.) *Getting started with schema.org using Microdata*. Retrieved 8 November, 2015 from https://schema.org/docs/gs.html [Archived by WebCite® at http://www.webcitation.org/6m45FdQtN]

Schwartz, C. (1998). Web search engines. *Journal of the Association for Information Science and Technology*, 49(11), 973-982.

Sulé, A. (2015) Schema.org, la mejora de la visualización de los resultados en los buscadores y mucho más. *BID: textos universitarios de biblioteconomía y documentación*, 34 (junio 2015). Retrieved 8 November, 2015 from http://bid.ub.edu/es/34/sule.htm [Archived by WebCite® at http://www.webcitation.org/6m45Q1K46]

Tallon, P.P. and Scannell, R. (2007) *Communications of the ACM*, vol. 50 (11), 65-69.

Trauth, E. M. The evolution of information resource management. *Information & Management*, vol. 16 (5), 257-268.

Wilson, T.D. (1989). Towards an information management curriculum. *Journal of Information Science*, 15, 203-210 Retrieved 8 Jun, 2016 from http://www.informationr.net/tdw/publ/papers/infmagt89.html [Archived by WebCite® at http://www.webcitation.org/6m45WWFPR]

Wilson, T.D. (2002) Information Management. *International Encyclopedia of Information and Library Science*, 2nd ed. Retrieved 18 November, 2015 from http://informationr.net/tdw/publ/papers/encyclopedia_entry.html [Archived by WebCite® at

http://www.webcitation.org/6m45eksAX]

Winter, A. F., Ammenwerth, E., Bott, O. J., Brigl, B., Buchauer, A., Gräber, S., ... & Heinrich, A. (2001). Strategic information management plans: the basis for systematic information management in hospitals. *International Journal of Medical Informatics*, 64(2), 99-109

Witehead, J. (1996) *Orality and Hypertext: An Interview with Ted Nelson.* Retrieved 12 July, 2015 from http://www.ics.uci.edu/~ejw/csr/nelson_pg.html [Archived by WebCite® at http://www.webcitation.org/6m45mdf2N]

Wright, J. (1997) The man who invented the web. *Time*, May 19, vol. 149 No. 20. Retrieved 18 November, 2014 from http://faculty.washington.edu/mfan/ebiz509/download/readings/TheManWhoInventedWeb_Time97.pdf [Archived by WebCite® at http://www.webcitation.org/6m45u1J7x]

Zhang, J. and Dimitroff, A. (2004) Internet search engines' response to metadata Dublin Core implementation. *Journal of Information Science*, 30 (4), 310-320. doi: 10.1177/0165551504045851