8-2010

Researcher Engagement with Web Archives: Challenges and Opportunities for Investment

Arthur Thomas

Eric T. Meyer

Meghan Dougherty
Loyola University Chicago, mdougherty@luc.edu

Charles van den Heuvel

Christine Madsen

See next page for additional authors

Recommended Citation

Thomas, A., Meyer, E.T., Dougherty, M., van den Heuvel, C., Madsen, C., Wyatt, S. (2010). Researcher Engagement with Web Archives: Challenges and Opportunities for Investment. London: JISC.

JISC

Researcher Engagement with Web Archives
Challenges and Opportunities for Investment

August 2010

Arthur Thomas
Eric T. Meyer
Meghan Dougherty
Charles van den Heuvel
Christine Madsen
Sally Wyatt

oiioiioi oxford
oiioiioi internet institute
oiioiioi university of oxford

Maastricht University

virtual knowledge studio

Acknowledgements

This report was funded by JISC, the Joint Information Systems Committee, from April to August 2010. The project was a partnership between the Oxford Internet Institute at the University of Oxford in the United Kingdom (http://www.oii.ox.ac.uk) and the Virtual Knowledge Studio at Maastricht University in the Netherlands (http://virtualknowledgestudio.nl/). Questions or queries about this report may be directed to:

Dr. Eric T. Meyer, Project Director
Oxford Internet Institute, University of Oxford
1 St Giles, Oxford, OX1 3JS, United Kingdom
Tel: +44 (0) 1865 287210
Email: eric.meyer@oii.ox.ac.uk

Neil Grindley, Programme Manager
JISC, Digital Preservation & Records Management
1st Floor Brettenham House (South), 5 Lancaster Place, London, WC2E 7EN, United Kingdom
Tel: +44 (0) 203 006 6059
Email: n.grindley@jisc.ac.uk

Please cite this report as:
Thomas, A., Meyer, E.T., Dougherty, M., van den Heuvel, C., Madsen, C., Wyatt, S. (2010). *Researcher Engagement with Web Archives: Challenges and Opportunities for Investment.* London: JISC

Table of Contents

Executive Summary

This report, which should be read in conjunction with its companion, "Researcher Engagement with Web Archives: State of the Art" (Dougherty, et al., 2010), looks beyond the current state of web archives, and the uses made of them, to expand on some of the challenges identified there, and to point out some of the important opportunities which exist for funding bodies to add considerable value to the investments made to date, as well as to move web archiving technology and practices to the next level of comprehensiveness and usefulness.

One of the biggest challenges is that the investment to date in web archiving has been woefully inadequate to allow technology and practice to keep up with the dizzying pace of innovation on the Internet. The introduction of new content formats (such as multi-media and dynamically executable content) has been accompanied by the evolution of completely new paradigms for content building and interaction, loosely grouped under the rubric Web 2.0, and particularly including user (and multi-user) generated content and the new social media platforms. All of these developments pose significant challenges to a web archive community which is still struggling to cope with Web 1.0 (that of largely static content). In addition, the so-called "deep web" – the web of contents which are hidden behind query interfaces, and derived on the fly from back-end databases – threatens to swamp the volume of traditional content, while posing unique problems of access to web archiving technologies built around crawlers. The most difficult challenge of all may be the apparent move of the web (and indeed of the internet as a whole) away from PC-centric, browser-based platforms towards a much more diverse set of user platforms (such as mobiles) which make use of "apps" (single-purpose, dedicated software) which often use proprietary communication protocols rather than HTTP (Anderson & Wolff, 2010).

These challenges carry with them new opportunities for innovation in technology and in practice. The significant opportunities which we identify in this report include:

- A move away from costly and time-consuming attempts to identify *a priori* the content (the "needles") likely to be of interest to web researchers, and towards what we call "collecting the haystacks": the rapidly-falling cost of storage, and new technologies and metadata conventions for managing multi-petabyte repositories, suggest that less effort should be placed on selection and collection strategies, and more on ways for users rapidly to survey, annotate, contextualise, and visualise those repositories, and to find and select the thematic elements of interest to them.

- Development of means to blur the distinction (in formats and naming conventions) between archival and live content. This would allow the web archive community to make use of the powerful search, annotation, visualisation and analysis tools developed for the live web, and to capitalise on the vastly larger investment which is available in the latter.

- The need for major investment in technical methods for collecting the new content types mentioned above. Ideally, this should be taken as an opportunity to move away from monolithic collection software to that based on web services, accompanied by development of workflow tools which put archivists and users on an equal footing in terms of building new thematic collections out of general-purpose ones. These tools would then allow virtual thematic collections to be considered as views (in the database sense) into the more general.

- The opportunity to capitalise on the rapidly-developing technologies and methodologies of the Semantic Web and Linked Data communities, especially the use of Resource Description Framework (RDF) and associated query and inference languages, for defining, representing, querying and integrating metadata about collections in much more flexible and powerful ways

than are possible with more traditional metadata conventions. These new methods would support much easier cross-linking of information from separate archives into meta-archives, thereby preventing wasteful duplication of collection efforts and storage, while giving users a much clearer picture of what content is available, and where.

- A major investment in overcoming the total functional inadequacy and poor scalability of existing web archive search and visualisation tools, preferably, as was mentioned, by blurring the distinction between live and archival content, while incorporating the new metadata methods to allow much more precise semantic search of large-scale archives.

- Encouraging the use of cloud storage architectures for archives, recognising their potential for significant economies of scale and for better data management practices.

- Integrating more closely with the Web Science community (Hendler, Shadbolt, Berners-Lee, & Weitzner, 2008) which is already exploring issues of how the web is developing, how content is created collaboratively by communities of interest, and how Future Internet architectures may overcome many of the issues (such as trust and privacy) with which the web archive community is currently struggling.

Introduction

The importance of the Internet for research, society, and the economy is unquestionable. However, content on the web and related social media is constantly in flux as it is updated, replaced, and deleted. Various efforts to archive the web or portions of it have been developed around the world. Much of this work has been done from the point of view of preservation for its own sake. Less work exists on how these preserved archives might then be used by researchers (those interested in the content for its own sake, as well as web scientists interested in the structure and dynamics of the web itself) and others to ask meaningful new questions.

This report is one of two aimed at starting to bridge this gap between archivists and researchers, and thus to build a compelling case for promoting, supporting and using web archives as a research resource. The first, entitled "Promoting Researcher Engagement with Web Archives: State of the Art," (Dougherty, et al., 2010) and aimed at archivists and researchers, summarizes existing work using archives and describes, using exemplar cases, the strategies developed for working with archives. The present report, intended for JISC and other funding bodies along with other interested parties, aims to identify current gaps in the funding for research using web archives and the development of appropriate tools for collecting and working with web archives. This analysis highlights opportunities where funding may help to remove barriers to progress by increasing the quality of tools and methods for researchers to exploit web archive resources. In this report:

- We analyse available tools and services, and identify the challenges to their effective use
- We suggest development initiatives that may boost researcher engagement.

This document should be read in conjunction with the above-mentioned "State of the Art" report.

Methodology

We undertook a combination of online research and structured interviews with stakeholders. These stakeholders were a mix of archivists, web archive users and web scientists; some of the people interviewed have multiple roles, which reflects the fact that many small archives are built by researchers for their own purposes. Appendix A lists the 17 people who participated.

Background

The web has become a vast, but often changing, and sometimes disappearing, storehouse of cultural, historical and scientific information. Several groups are now successfully archiving large portions or selected segments of the web. Through these activities, they aim to create an archival record of web culture or of contemporary culture as manifested on the web. This record is intended "to resemble a digital library" from which historians, curators and scholars can draw data to support their research (Lyman & Varian, 2003).

The traditional practices of the field of library and information science have come to dominate web archiving. They are a good fit because the practices built into these fields are technologically well-developed and ready to handle the content delivery systems required by web archives. Further, they offer an existing policy framework for the collection of contemporary cultural materials. But, there are consequences to the situation of web archiving within libraries and archives. Library and information science practices have influenced the development of web archives regardless of why those archives were developed or how they will be used. This has set up a point of contention between librarians and information scientists, who would like to build widely valuable and accessible collections, and humanities and social science researchers who would like to use web archives as a

basis for understanding digital cultural heritage or web historiography. The two perspectives are not diametrically opposed, but there are certainly points of conflict that are derived from differently held philosophical undercurrents that motivate each.

As a result, large libraries and archives continue with their efforts to build large multi-purpose web archives, while researchers - either on their own, or partnering with archivists - develop their own archives for use in their research. Archival institutions find it difficult to support the development of focused, project-specific archives, but researchers cannot yet find value for their work in the large multi-purpose archives being built by archivists. The core tools for creating basic web archives are now widely in use, but there is no underlying infrastructure in place to support the research into these archives.

Current web archiving projects and initiatives fall into three categories:

- **Large-scale collections**, led by the Internet Archive (the largest), have to date focused on collecting rather than on use. These initiatives take a whole-domain approach focusing on archiving as much of the public web as possible.

- **Researcher-led initiatives**, characterized by small, purpose-built collections. These collections are created within the scope of a particular research question and seldom used by anyone other than the researcher.

- **Institution-led initiatives** and **consortial efforts** are selective, thematic, deposit-based or a combination of these approaches. These web archiving initiatives are often intended to extend the existing collection development policies of cultural institutions on the web.

With the exception of the large-scale archives such as the Internet Archive (IA), which maintain a commitment to serving the 'general public,' most of these initiatives are targeted at creating collections of use to researchers. The current gaps in understanding regarding use of web archives lies precisely at the point where the two meet. That is, we know some of what researchers want from web archives and we are just now beginning to know what web archives can provide, but web archives of use to researchers have not been available long enough yet to develop a comprehensive understanding of how researchers are engaging with web archives.

Not much is known about users' behaviour in web archives. Most archiving institutions therefore rely on semi-hypothetical use cases to refine and expand their usability and interfaces. One particularly detailed study was conducted at the National Library of the Netherlands (Ras & van Bussel, 2007). This structured experiment, run similarly to a task-oriented usability study, evaluated user comfort level with search and access tools and attempted to determine user satisfaction with archive contents. Several use-scenarios were posited. Few native users have been studied to date, and reports of these studies remain unpublished works in progress. We do not have much to draw on when speculating about users in web archives. However, those who are developing their own web archives for directed and narrow research purposes can provide some insight about how they use their archives to produce knowledge in their field.

Challenges and Opportunities

In this report, we endeavour to identify the challenges currently facing the web archive community, both archivists and users, and the opportunities available to funding agencies to help the community meet those challenges. For the sake of analysis, we distinguish between challenges facing archivists, users and web scientists, but clearly there is substantial overlap and complementarities between them. In particular, overcoming some of the archiving challenges will make archive contents a good deal more attractive, reliable, comprehensive and useful to users. We believe that these are all areas where clear-sighted and well thought-out funding initiatives can have a huge positive impact on the quality and usefulness of web archives, and can foster new approaches which address the inherent conflicts of goals and means that were described above. For each challenge, we aim to identify the concrete opportunities for funding agencies to tackle it.

Challenges to Archiving

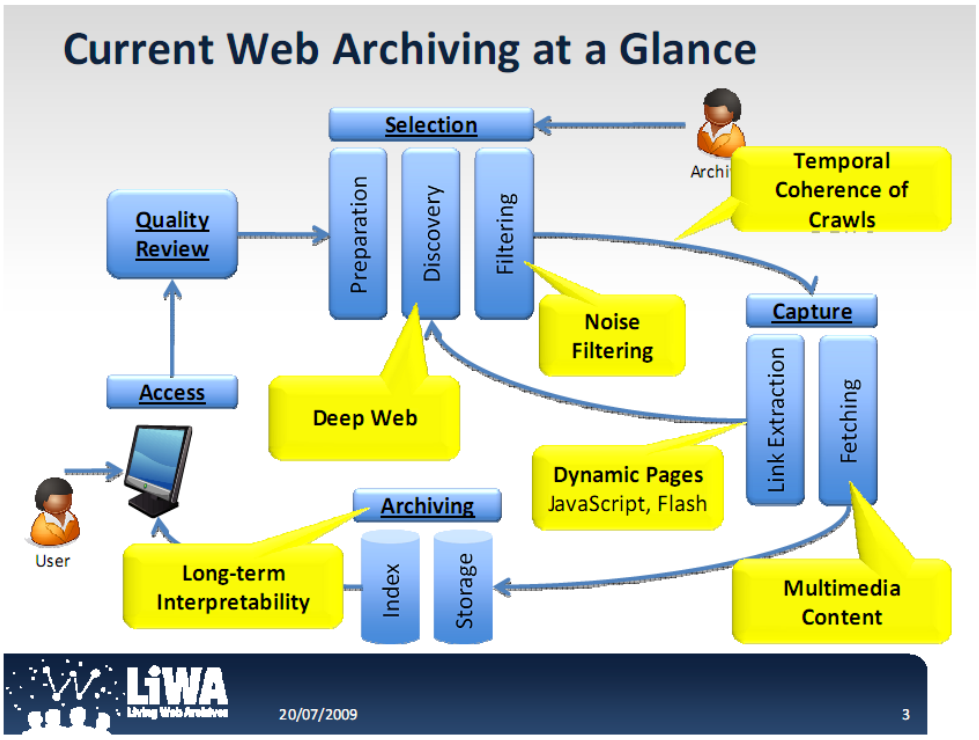

Figure 1: Overall workflow of web archiving (Reproduced from Risse, 2009)

Figure 1 summarises the processes involved in web content capture, and highlights some of the challenges which yet remain to be overcome.

Content Selection: approaches and strategies

Deciding *what* should be collected is the first and often the most difficult challenge facing both archivists and users. Traditionally, a distinction has been made between *selective harvesting* and *domain harvesting,* with the former typically being undertaken by small academic groups, and the latter being the purview of global collection institutions such as IA and the various national archives and libraries.

Defining a collection strategy has typically been a manual process, often not well-documented and hence often inconsistently applied. It is clear that better tools are needed to allow the person

defining a collection strategy to gain a meaningful overview of the target content, to ensure that only content relevant to the collection's purpose is actually collected.

Another problem which is growing in importance is that web content is often created dynamically, either in response to a user-defined context (e.g. existence of cookies, use of Javascript for dynamic page creation, dynamically-imbedded advertisements, etc.) or via forms-based interfaces which create content as a result of a search of an underlying database (the so-called "deep web"). In each of these cases it is difficult to identify exactly what the target content may be, or how it will be rendered, thereby making the selection process more hit and miss. A distinction can be made between capturing the *content* and capturing *the way that the content is experienced*. An archiving strategy must decide whether to capture only content, or to attempt the much more difficult task of capturing the appearance(s) and behaviour(s) in all their possible varieties. These hard problems are discussed in more detail below.

To date, most harvesting has been implemented by means of *link crawling*, i.e. recursively following embedded hyperlinks to some depth. This form of harvesting is fraught with technical and legal difficulties. The technical problems arise from broken links, circularity, high noise content such as advertisements, incomplete "robots.txt" exclusion files and similar artefacts. The legal problems arise because of difficulties in establishing clear permissions to copy and/or publish content. One approach to overcome these problems would be to make much more systematic use of Sitemaps.[1] These are XML files, created by the owner of a web site, describing the accessible content of a site, together with descriptive metadata including explicit permission grants and heuristics, such as timestamps and priority information, for guiding collection along lines laid down explicitly by the owner, who presumably has the best understanding of the structure and semantics of the content graph.

Opportunity: Support development of better tools to allow archivists to review Sitemaps and other structural overview visualisations, and to incorporate them into collection workflows. Encourage Web owners to create Sitemaps, and support development of collection tools which can use Sitemap information.

Content Capture: Tools and Workflows

Having decided what to collect, the problem remains of *how* efficiently and consistently to implement the collection strategy.

There is still little consensus on collection practices or the use of specific tools, although international collaboration in the field has led to some convergence on certain crawlers (primarily the Internet Archive's Heritrix[2]) and the development of the new WARC[3] file format as an international standard (ISO 28500: 2009), by the IIPC (International Internet Preservation Consortium) for content. However, as discussed below, these collection methods and formats are still not adequate to deal with rapidly evolving web formats and complex structures.

The last few years have seen the development of a few widely used integrated tool suites, aimed primarily at archivists, (providing some workflow management, crawler integration, permissions management, quality monitoring and some support for metadata). These include:

[1] http://www.sitemaps.org/
[2] http://crawler.archive.org/
[3] http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

- NetArchiveSuite[4] (developed by the two Danish national deposit libraries) is a tool for planning, scheduling and running web harvests. It supports small, thematic harvests (e.g. related to special events, or special domains) as well as harvests of entire national domains. It has good functionality for quality monitoring. It is distinguished by being fully distributable across networks of machines.

- Web Curator Tool[5] (WCT, developed by the British Library and National Library of New Zealand, and now maintained by Oakleigh Consulting Ltd.) is an open-source workflow management application for selective web archiving. WCT is designed for use by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the Heritrix crawler and supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata. Filters can be defined to include or exclude content.

- The Pandora Digital Archiving System (PANDAS[6], developed by the National Library of Australia), which uses HTTrack[7] as its crawler. Its filtering capabilities are quite weak. One distinguishing feature of PANDAS is that it assigns a system generated running number to each title when it is registered. This number becomes part of the *persistent URL* applicable to each archived title's entry page.

- The OCLC Web Archives Workbench[8] suite of web archiving tools which emphasises management of archived content as *aggregates* rather than as individual objects.

- The Archive-It[9] subscription service from the Internet Archive, which allows institutions to build and preserve collections of digital content. Archive-It partners can harvest, catalogue, manage, and browse their archived collections. Collections are hosted at the Internet Archive data centre and are accessible to the public with full-text search.

- The Living Web Archive project (LiWA)[10] is a relatively young, but important, initiative (first technology release in June 2010) with a primary emphasis on methods and tools for capturing rich (multimedia) content, and on archive quality improvement, especially by noise reduction (removal of spam and other unwanted content) and by improving the temporal coherence of collections (i.e. ensuring that all content is captured within the same timeframe) (see Risse, 2009).

- Taverna[11] is not a web archiving tool *per se*, but is rather a general purpose workflow management platform. Taverna has a large developer community, and offers the possibility of providing ways to share workflows, which are documented and reusable sets of steps and transformations, to better support reusable and transportable web archives.

[4] http://netarchive.dk/suite/
[5] http://webcurator.sourceforge.net/
[6] http://pandora.nla.gov.au/pandas.html
[7] http://www.httrack.com/
[8] http://sourceforge.net/projects/webarchivwkbnch/
[9] http://www.archive-it.org/
[10] http://www.liwa-project.eu/
[11] http://www.taverna.org.uk/

Opportunity: All of these tools have been built as more or less stand-alone, monolithic, systems, with little possibility for re-use of components across different platforms. It would be worthwhile to consider **re-implementation of these tools within a Web services paradigm, which would allow greater re-use, commonality and flexibility to modify tools for specific needs**. One opportunity which seems to have been overlooked is the **use of public-domain workflow management systems such as Taverna for composing complex workflows out of Web services.**

Wrappers and Digital Bundling

> *The discrete 'item' — the book, the journal article — is becoming less and less relevant in today's interconnected world. [...] For the library, what this means is that collections work will gradually need to shift from a focus on discrete items, to a focus on comprehensive collections and links both within and outside of collections.* (Morrison, 2007)

The WARC archive file format (ISO 28500:2009) is now becoming widely used; it extends the original ARC format by supporting recording of HTTP request headers, arbitrary metadata, duplicate management, etc. However, WARC is just a specification for a *container*, and says nothing about the formats or semantics of the objects contained within WARC files, and nothing about their relationships to each other.

There are a variety of relatively immature ongoing efforts to develop XML-based formats to bundle all of the content and metadata for a digital object in a more structured package; examples include FOXML[12], METS[13] (discussed in more detail below), MPEG-21[14] DIDL[15], MXF[16] and XFDU[17]. But a major obstacle to the use of these standards is that the Open Archival Information System Reference Model (OAIS[18]), which is intended to provide a framework for the use of such bundle specifications, although given widespread lip-service, has not been widely adopted, so that all of these efforts lack coherence. The Open Annotation Consortium[19] is also using the Linked Data paradigm (Hunter, Cole, Sanderson, & Ven de Sompel, 2010), and represents an opportunity to deal with the important ability to annotate objects in web archives.

Opportunity: Promote wider adoption of OAIS and OAC models. The use of the Linked Data paradigm (discussed in more detail below) to represent metadata would also make it a good deal easier to deal with this issue, since multiple structural relationships, rather than a strictly hierarchical view, could be better represented.

Improving Archive Fidelity and Quality: Noise Filtering and Temporal Coherence

Validation of captured content, i.e. ensuring that the content captured is in fact that which was selected, that nothing is missing, and that extraneous material, however that is defined based on the purpose of the web archive, is excluded, is a significant challenge. Currently, this is largely achieved through tedious manual review of crawler logs and the captured content. Better tools are needed for this important function.

[12] http://www.fedora-commons.org/download/2.0/userdocs/digitalobjects/introFOXML.html
[13] http://www.loc.gov/standards/mets/
[14] http://mpeg.chiariglione.org/standards/mpeg-21/mpeg-21.htm
[15] http://mpeg.chiariglione.org/standards/mpeg-21/mpeg-21.htm#_Toc23297974
[16] http://en.wikipedia.org/wiki/Material_Exchange_Format
[17] http://sindbad.gsfc.nasa.gov/xfdu/pdfdocs/xfduspec.pdf
[18] http://public.ccsds.org/publications/archive/650x0b1.pdf
[19] http://www.openannotation.org/

Spam and advertising removal depends on understanding what the principal focus of a collection ought to be, and detecting content which is anomalous with respect to that focus. And of course, in some contexts the spam and advertising may have intrinsic interest and value to archival researchers, so should not be removed at all! There is a continuing arms race between spammers and search engine developers (Benczúr, et al., 2008), but little of this expertise seems to have percolated into the web archive community. For example, spammers make use of *cloaking* which uses the web's inherent content negotiation to detect whether a request originates from a browser or from a crawler. Linkage analysis and statistical analysis of page contents can aid in filtering. A variety of computational linguistics techniques e.g. using unsupervised learning (Guthrie, 2008; Guthrie, Guthrie, Allison, & Wilks, 2007) have been applied to detect anomalous content. Benczúr et al. (2008) also propose the development of a shared blacklist of spam pages. Increasingly, trust-based methods (Guha, Kumar, Raghavan, & Tomkins, 2004) are also being applied.

Temporal coherence may be lost when site content is changing rapidly compared to the crawl time. Spaniol, Mazeika, Denev, & Weikum (2009) discuss methods that arise for detecting, measuring and repairing coherence defects. They present visualisation strategies, such as that shown in Figure 2, which can be applied on different level of granularity to detect such defects.

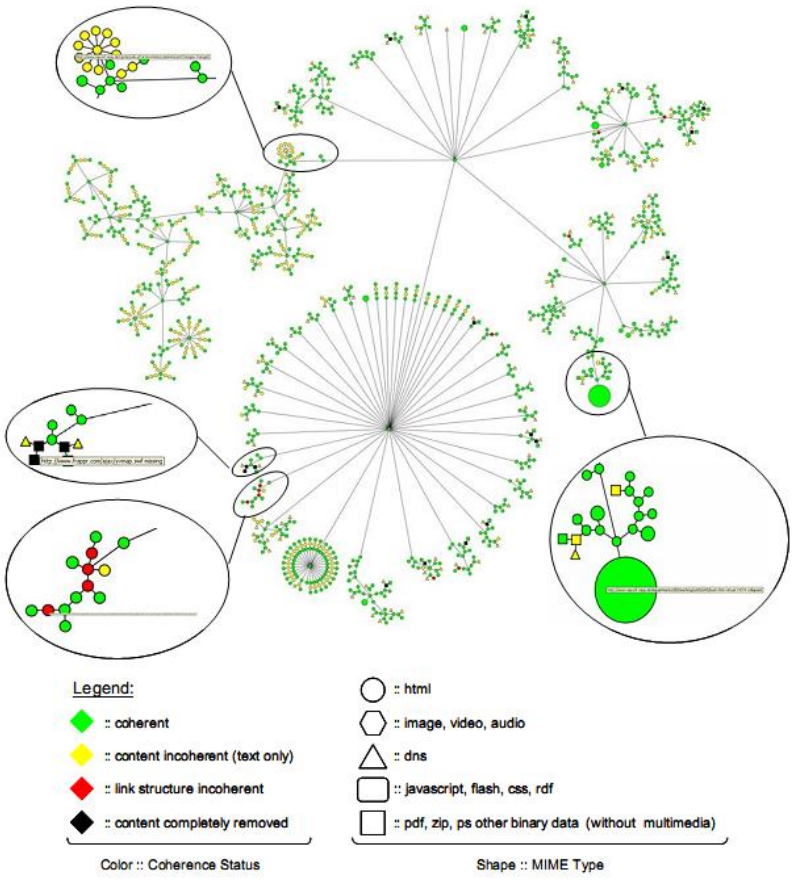

Figure 2: Visualisation of coherence defects
(Reproduced from Spaniol, et al., 2009)

Opportunity: Support and extend projects (such as LiWA) which begin to address collection of rich content, improved collection quality and suppression of noise content.

Reducing Overlap between Archives

Historically, there has been considerable overlap and duplication between the collections of different archives. While this may be good from a redundancy point of view, it is very wasteful of resources such as bandwidth and storage. In addition, users can find overlapping collections confusing and time-wasting. The problem is that there is presently no adequate mechanism for archives to publish summary descriptions of their content; this problem is being addressed by initiatives such as Archives Hub (discussed below).

Opportunity: Create and support national and international forums for planning collections to avoid excessive overlap or duplication, and continue to support community portals such as Archives Hub which allow archivists and users to determine what has been collected and where.

Adequacy of Metadata

Metadata in this context is information which enables and documents the long-term preservation of and access to digital objects. At a minimum, metadata maintained by archives should include:

1. *Provenance*, describing the custodial history of the object
2. *Authenticity*, validating that the object is what it purports to, and has not been modified
3. *Preservation activity*, describing actions taken to preserve the object
4. *Technical environment*, describing the IT environment necessary to render the object faithfully
5. *Rights management*, recording any property rights which may govern retention or publication of the object.

(Lavoie & Gartner, 2005)

There is a plethora of sometimes conflicting and overlapping standards for archival metadata (see Riley, 2010 for a good overview). The most ambitious effort to date is the Open Archival Information System (OAIS[20], ISO 14721:2003) reference model (CCSDS, 2002), which defines concepts and responsibilities essential for preservation of digital information. Like many such reference models, and because of its complexity, it has been adopted only in parts by archival institutions. For a comprehensive evaluation within the UK context, see Allison (2006).

More recently, and perhaps with a greater likelihood of being widely adopted, the Metadata Encoding and Transmission Standard (METS[21]), developed by the Library of Congress, is an "*XML Schema designed for the purpose of creating XML document instances that express the hierarchical structure of digital library objects, the names and locations of the files that comprise those objects, and the associated metadata*." The METS standard, while supported by a wide variety of open source and commercial tools, suffers, like many standards derived from the library community, and especially from the MARC tradition, from an overly restrictive view of born-digital object collections, and it has thus proved difficult to adapt it to web archiving needs (Guenther & Myrick, 2006).

An important recent development is the potential for using Linked Data[22] approaches to publishing, integrating and searching metadata. This is an alternative to the rigid hierarchical approach of record-based standards such as METS. In the Linked Data approach, all objects and relationships are defined by a labelled, directed graph (which can be *serialised* in a variety of RDF[23] syntaxes). In such

[20] http://public.ccsds.org/publications/archive/650x0b1.pdf
[21] http://www.loc.gov/standards/mets/
[22] http://linkeddata.org/
[23] http://www.w3.org/RDF/

a graph, unlike hierarchical models, there are no distinguished nodes, and users can enter the graph at any node and then browse by edge following. This provides a great deal of flexibility and enables users to view the underlying data in a very flexible way that matches their needs.

Each of the objects and relations in an RDF graph is represented by an URI[24], but the issue arises of how to use persistent naming schemes. The Persistent URL[25] approach can be used to solve this problem, by providing a layer of indirection, but considerable work is required to develop URL naming schemes for particular domains (see the discussion on naming below). The upside is that there is now a huge effort within the Semantic Web community to develop tools for managing large RDF graphs, and it also leads naturally into the use of reasoning languages such as OWL for making inferences over these graphs.

Opportunity: Support development of metadata standards and tools built around the Linked Data model, as an alternative to record-based standards such as METS.

The Deep Web

One of the emerging challenges to traditional crawler-based web archiving is that more and more content is not easily accessible to crawlers, but is instead is hidden behind web forms and other kinds of query interfaces (Bergman, 2001; Wright, 2009). Clearly the volume of information represented by this deep web is potentially many orders of magnitude larger than the static surface web.

There is currently intensive work under way to develop methods for retrieving deep web content. For example, DeepPeep[26] (Barbosa & Freire, 2007) is a search engine which retrieves data via keyword-based interfaces and forms. It attempts to classify web forms according to their conceptual domain (e.g. used cars) and builds a set of potential form entries which can be used to elicit retrieved content. Google and other search engines have supported the use of Sitemaps to allow webmasters to create searchable descriptions of content which is not reachable by crawling. Google's fledgling Deep Web system pre-computes submissions for each HTML form and adds the resulting pages into the Google index. The pre-computing of submissions is done using three algorithms: (1) selecting input values for text search inputs that accept keywords; (2) identifying inputs which accept only values of a specific type (e.g., date); and (3) selecting a small number of input combinations that generate URLs suitable for inclusion into the web search index. In a different approach, the Apache Web Server *mod_oai* module (Nelson, Smith, Del Campo, Van de Sompel, & Liu, 2006) makes server content accessible through the OAI-PMH[27] protocol which is widely used within the digital library community. However, very few of these approaches seem yet to have been taken up by the web archive community.

One tool that has been developed for archiving objects from database-driven websites is DeepArc[28] (Bibliothèque Nationale de France). Users use a form-based search interface to enter keywords which are used to query the database. DeepArc requires access to the underlying database schema, and creates a map to its own target data model.

[24] http://en.wikipedia.org/wiki/Uniform_Resource_Identifier
[25] http://www.purl.org/docs/index.html
[26] http://www.deeppeep.org/
[27] http://www.openarchives.org/OAI/openarchivesprotocol.html
[28] http://bibnum.bnf.fr/downloads/deeparc/

Opportunity: Gaining access to Deep Web content is one of the greatest challenges to present day archiving. Greater priority should be given to methods for addressing this problem, and especially to methods for understanding the underlying data schemas and semantics of form interfaces.

The Dark Web

The Internet, unlike the web, supports many closed online communities (sometimes using non-standard or proprietary protocols). Important examples include virtual worlds and massively multiplayer interactive games such as Second Life[29] (SL) and World of Warcraft[30] (Antonescu, Guttenbrunner, & Rauber, 2009; PVW, 2010). According to Linden Labs (Linden, 2010), the developer of Second Life, a total of 481 million user hours were spent on SL during 2009, with 769,000 monthly repeat logins and it is growing at approximately 20% per annum. In financial terms, the SL virtual economy had a GDP of $567 million dollars, a growth of more than 65% in one year.

In such virtual worlds, there is not only a need to capture content, but capturing as well *the way that the content is experienced,* by documenting *patterns of interactions.* Current archive formats do not provide suitable ways of documenting interactive fiction and games at the bit-level: they do not provide the means to interpret the raw bits as higher-level data constructs. In the case of highly complex, interactive objects such as hypertext fiction and games, inadequate representation information severely hampers preservation of these works. The Preserving Virtual Worlds project[31] is defining metadata standards aimed specifically at archiving virtual worlds. This will require:

- development of new schema to capture technical metadata and other representation information for the data formats included in our case studies;
- a new schema for description of context information for digital objects;
- a new schema for preserving complex interactive user-behaviour;
- a new schema for structural metadata to encode interactive fiction;
- a set of suggested elaborations of existing wrapper formats to allow for complete support of representation information.

  (PVW, 2010)

They propose the inclusion of representation information through a revised version of the METS format and/or development of the Electronic Literature Organization's (ELO) proposed X-Lit[32] format.

Opportunity: Capturing virtual worlds and other interactive contexts is becoming important, since, along with social media (see below) they represent a huge set of content which is not at all addressed by current archiving efforts. The work of projects such as PVW should be supported and extended.

[29] http://secondlife.com/
[30] http://www.worldofwarcraft.com/index.xml
[31] http://pvw.illinois.edu/pvw/
[32] http://www.eliterature.org/pad/bab.html

Stateful Transactions and Executable Content

The existence of state (e.g. represented by *cookies*) can alter the apparent content of web pages in ways that are highly context dependent, and thus difficult for web crawlers to emulate. Even more challenging, widespread use of browser-side scripting languages (CSS, Java/ECMAscript, ActiveX, AJAX, Flash, Silverlight, Java applets, etc.) can drastically modify the apparent behaviour of a web page and thus the overall user experience, and can introduce large quantities of state. Browser plug-ins such as AdBlock[33] or NoScript[34] can also change the overall behaviour in ways that can be controlled at fine grain by the user.

HTML5[35] will eventually simplify some of these issues, by providing a more coherent and standardised framework for plug-ins and codecs, but is unlikely to be widely adopted before 2012 at the earliest.

Content Negotiation and Transcoding

From its earliest days, the HTTP specification supported the concept of content negotiation which makes it possible to serve different versions of a web resource, depending on the preferences of the user agent being used to retrieve it. One implication of this is that a crawler may retrieve content in a different format from a browser. More recently, the introduction of *transcoding* technology supports on-the-fly changes in format. These techniques are becoming much more widely used as a consequence of the wide-spread adoption of mobile browser platforms, which may have reduced capabilities.

The impact of this development on archiving is significant. No longer can a crawler be relied upon to retrieve the same content, in the same format, as a user would see, especially using a non-PC-based browser. To date, very little work has been done in the web archive community to address this problem.

Opportunity: Support a much more aggressive approach to capturing variable content presentation, in conjunction with work described above to handle dynamic and executable content.

Mobile Content, Apps and Servers

As mentioned, current archiving approaches take a predominantly PC-centric view of the web – they assume that content will be viewed using a PC-based browser. But increasingly, mobile platforms such as smart phones are becoming the primary viewing platforms, and transcoding mechanisms are often used to dynamically re-render content for multiple platforms. We will need a wide range of emulation mechanisms to capture all of this variety for historical viewing.

An even greater challenge is the use of smart phone (especially iPhone and Android apps) as dedicated mobile platforms, using protocols other than HTTP (many of them proprietary and not publicly-documented) for content exchange. How are archivists to deal with this plethora of new closed content?

As the Web of Things (physical objects which have associated web sites that advertise their content or status) grows ever larger, devices are also web *servers* as well as clients – they serve up content

[33] http://adblockplus.org/en/
[34] http://noscript.net/
[35] http://www.w3.org/TR/html5/

about their location, etc. An interesting example of technology to support this idea for mobile devices is Opera Unite[36], which supports serving local content from within a web browser.

Opportunity: The development of better emulation mechanisms will allow the web archiving community to move away from a PC-centric view of web archiving towards one which copes with content generated by, and viewed on, a wide variety of mobile and embedded platforms.

Multimedia Content

Currently, web archiving is notable for its lack of systematic methods for collecting, tagging and searching the huge amounts of multi-media content, such as images and videos, which are becoming an important domain of web content.

One early example of an approach to this problem is LiWA[37] (whose first product is a plug-in for Heritrix (LiWA 2010) to capture YouTube videos). Another is TubeKit[38], a toolkit for creating YouTube[39] crawlers. A big challenge is that the formats of sites such as YouTube change every few weeks or months.

Of course collecting is only one problem; the other is searching such content once it is collected. Currently tools for searching multimedia content are quite primitive.

Opportunity: Support development of tools for capture and search of multimedia content.

Web 2.0 and Social Media: Streaming Content

The rapid recent development of user-editable content (captured under the umbrella term Web 2.0) poses huge challenges for web archiving. One particularly problematic aspect is that the growing set of dynamic Internet sources, including rapidly changing collaborative content such as Wikis and social media such as Twitter[40] and Facebook[41], can create real-time streams of data, often containing valuable information about public sentiment and the Zeitgeist. They change on a much shorter timescale than typical crawl times, so alternative approaches are needed. Some such sources make historical dumps available to selected consumers, while others provide some sort of Application Programming Interface (API). In the latter case, the issue of scale become paramount, since consumers may have to deal with data volumes measured in millions of updates per day.

Wikis

Wikis have become an important store of community knowledge, and support scholarly collaboration in a wide variety of disciplines. There are a large number of available Wiki platforms (MediaWiki[42] being one of the most widely-used) and each presents slightly different challenges to would-be archivists. Several approaches to archiving seem possible:

[36] http://unite.opera.com/
[37] http://www.liwa-project.eu/
[38] http://www.tubekit.org/download.php
[39] http://www.youtube.com/
[40] http://twitter.com/
[41] http://www.facebook.com/
[42] http://www.mediawiki.org/wiki/MediaWiki

1. Acquire frequent snapshots, but the definition of frequent may need to vary widely, from minutes to months, depending on the rate of change of a wiki's content.

2. Capture the public changelogs which are maintained by all the important platforms. These can be very large in volume (the largest public Wiki, the English Wikipedia, currently averages 130,000 edits/day, with spikes as high as 300,000, and of course Wikipedia is available in 90+ additional languages), so reconstructing the underlying text at any instant in time may require significant computation.

3. Acquire periodic dumps from the wiki's owner. An interesting recent innovation is the availability of (a structured subset of) Wikipedia's content in the form of a set of RDF triples (e.g. DBpedia[43], whose most recent version contains information about 3.4 million things, many of them classified using a consistent ontology, amounting to a total of over 1 billion RDF triples. Currently DBpedia releases are made approximately every 6 months).

Twitter

The Twitter microblogging service has grown in just four years to be a significant source of Internet content, but there is considerable confusion and disagreement within the web archive community about whether this content is sufficiently interesting to be worth collecting *en masse*. Recent adoption of Twitter by large-scale news sources and other commercial organisations suggests that it will become a valuable historical repository, worthy of collection. And the recent introduction by Twitter of a low-latency, high-volume streaming API[44] (the so-called "Firehose") opens up the possibility of archiving and searching significant subsets of tweets in close to real-time, representing a current data flow of up to 55 million tweets per day. The elements of this data stream are returned in either XML or JSON[45] formats. Currently, the full-volume Firehose is only available to partners such as Google, Microsoft and Yahoo! and selected commercial start-ups, but it is expected that it will become much more widely available within a short time. Innovative user interfaces such as Google Replay[46] enable display and query of timelines of the Twitter feed. Much of Twitter's content is not created or viewed through the Twitter.com web site, but through any of a very large number of third-party apps (developed for both PC and mobile platforms). It is clearly going to be very difficult if not impossible for searches to capture the content as rendered via these platforms.

In April 2010, the US Library of Congress announced that it would begin to archive public tweets (retrospectively to March 2006). This archive will be available with a 6 months delay after collection.

Facebook

Facebook is accessible only to a small, approved list of crawlers, entries are password-protected and Facebook's Terms of Use prohibit "data mining, robots, scraping or similar data gathering or extraction methods." But the need for archiving becomes even more relevant when it is remembered that while Facebook now has over 500 million users, (1) it does not guarantee to maintain information forever, and (2) many other important Web 2.0 sites have already disappeared, taking their users' information irretrievably with them.

The Facebook API[47] suite makes available a number of different (but not all) aspects of the Facebook collection of objects, albeit bound about with various security constraints. The Social Graph API

[43] http://blog.dbpedia.org/
[44] http://dev.twitter.com/pages/streaming_api_concepts
[45] http://www.json.org/
[46] http://googleblog.blogspot.com/2010/04/replay-it-google-search-across-twitter.html
[47] http://developers.facebook.com/docs/

enables access to the set of objects and social relationships that are defined by users' pages. Various search APIs and an SQL-like relational query language, FQL, are also available.

Facebook content presents significant challenges to archiving (McCown & Nelson, 2009). Even assuming that the APIs give adequate access, the complex and ever-changing terms of use, permissions policies and individual privacy preferences make archiving a considerable, even well-nigh impossible challenge (for a more extensive discussion of the privacy issues here, see the "Intellectual Property and Related Legal Issues" section below). And, as with many social media, capturing the appearances and behaviours may well be beyond current archiving techniques.

Social Media Consolidated Feeds/Analysis

A number of commercial companies now provide high-bandwidth feeds from various social media sites, and tools for filtering and analysing those data, including:

- Spinn3r[48], which provides a real-time feed of up 1 million posts/hour from 40 million blogs, 10 thousand mainstream news sources and more than 30 social media sites; feeds can be filtered by author, language, tags, link count and other criteria

- Visible Technologies[49], which crawls over half a million Web 2.0 sites a day, scraping more than a million posts and conversations taking place on blogs, online forums, Flickr, YouTube, Twitter and Amazon (but not Facebook at present). Customers get customized, real-time feeds of what's being said on these sites, based on a series of keywords.

- Recorded Future[50], which aggregates postings from a wide variety of social media sites and blogs, and applies a proprietary "temporal analysis engine" to extract important entities, events and associated timestamps, and to calculate sentiment and momentum. They also provide a suite of visualisation tools for exploring temporal changes and linkage graphs.

Another interesting recent approach is the Google OpenSocial API[51], which defines a common API for social applications across multiple Web 2.0 platforms. With standard JavaScript and HTML, developers can create apps that access a social network's friends and update feeds. This API is currently supported by, among others, MySpace, Friendster, Yahoo!, Plaxo and LinkedIn, but not by Facebook.

Opportunity: Support development of tools for capture and analysis of Web 2.0 sources. These could include archival browser simulators inspired by early proposals such as Browser Monkey (Tofel & Vahlis, 2006), which do a better job of capturing interactive behaviour. Also promote development and adoption of the Google OpenSocial or similar APIs.

Semantic Web and Linked Data Universe

The Semantic Web (Berners-Lee, Hendler, & Lassila, 2001) (sometimes also called Web 3.0) is shorthand for a related group of data and metadata models and tools (including reasoning tools) built around the W3C Resource Description Framework (RDF). RDF has come to be used as a general method for conceptual description or modelling of information that is implemented as web resources, using a variety of syntax formats. The overall goal of the W3C's Semantic Web activity

[48] http://spinn3r.com/
[49] http://www.visibletechnologies.com/
[50] https://www.recordedfuture.com/
[51] http://en.wikipedia.org/wiki/OpenSocial

(Berners-Lee, et al., 2001; Herman, 2010) is to promote creation of a Web of Data (in parallel to, and integrated with the Web of Documents), and thereby to enable machine navigation of the web and the use of semantic technologies to improve data representation, retrieval and reasoning. The closely-allied Linked Data[52] project aims to link together information from many different web repositories (using dereferenceable URIs, many in RDF) in a way that facilitates data integration. The current so-called Linked Data Universe consists of approximately 13 billion triples, with many millions of links between them.

The Semantic Web and Linked Data pose a whole new set of challenges for the web archiving community. First, they introduce a major new family of metadata standards which need to be incorporated into existing tools and methodologies. Second, the new data formats present challenges to archives built on traditional, file-based text formats; the RDF triple format presents much more structure. Third, the rapidly-growing Linked Data Universe, and its links to other repositories, forms a hug resource with which the archiving community must cope.

Opportunity: Promote the use of the Semantic Web paradigm in the traditional web archiving community, and support integration of the Semantic Web and Linked Data data models into existing web archive tools. Promote the use of triple stores as repositories for archive metadata (and possibly also data when appropriate), and the use of Linked Data browsing tools by archive users.

Archive Storage in the Cloud(s)

Traditionally, the cost of data storage has forced the archiving community to think long and hard about what it wants to archive. But the world of data storage is changing ever more rapidly. The cost of storage technologies (measured in £/GB) is halving approximately every 12 months, and seems set to continue to do so for the foreseeable future. Additionally, the arrival of cloud storage vendors, such as Amazon[53] and Rackspace[54], promises another inflection point in the cost of storage, as well as a huge degree of flexibility to grow storage resources on an as needed basis. Taken together, these trends suggest that it is now becoming economical for web archives to take a much less selective ("collect the haystack, not the needles") approach. This approach may not only be more cost-effective, reducing as it does the up-front cost of collection selection, but also permits much more flexibility for users to view and search collections using criteria which were not defined ahead of the time when the archive was created.

Opportunity: Analyse and promote the use of cloud storage, and encourage a change from a highly-selective to a much more sweeping approach to building collections, while recognising that this dictates the need for much better search and navigation tools, to allow users to find the needles they need in the much larger haystacks.

Intellectual Property and related Legal Issues

Recent developments in copyright legislation in the UK have the potential to completely change the landscape of web archiving as well as the landscape of traditional archives. Current law prohibits any UK library from archiving sites without permission from content creators. The 2003 *Legal Deposit Libraries Act*, if extended, could change this drastically, requiring that all forms of non-print publication comply with the *Copyright Act* and be legally deposited in a library. This is an important

[52] http://linkeddata.org/
[53] http://aws.amazon.com/ebs/
[54] http://www.rackspacecloud.com/cloud_hosting_products/files

step in preserving contemporary cultural heritage materials, but will put sudden and immediate emphasis on the need to make web archives more usable and accessible to scholars. Of course, any archive collecting from sites outside the UK must take note of applicable copyright laws of those jurisdictions. While the *Berne Convention for the Protection of Literary and Artistic Works Article 7(1)* requires that signatories grant copyright protection for the life of the author plus 50 years after his or her death, some jurisdictions, such as the European Union and the United States, define the term as "life +70 years". Understanding and implementing those constraints poses a major challenge for archive collection.

There are other significant legal issues which also potentially have a large impact on the collection and use of web content:

- *Data protection laws* specify what uses can be made of personally-identifiable information, and how, where and how long such information can be retained. Current archiving practices and tools do not easily support such constraints.

- *Freedom of Information* laws lay out conditions under which content created by public bodies must be made public, or exempted. Archives will need to respect these conditions.

- Increasingly stringent laws on creation and storage of *offensive or obscene content* expose archives to liability, even when such content is collected inadvertently.

- Social media sites such as Facebook have ever-changing *terms and conditions for their use*, and any archives of those sites could potentially be required to observe those constraints. Could this mean, for example, that content should be removed from archives when a user deletes her Facebook account?

- Many countries (China, Australia, and to a lesser extent the UK and others) have begun to impose increasingly onerous constraints (such as firewalls) on what content may legally be shown within their national boundaries. Does this mean that archives should (or must) also respect those constraints? If so, then they will have to tackle the problem of how to make content selectively accessible, without (as has been the case on several occasions recently) excluding entire domains (e.g. Wikipedia.org) from being accessible. One approach to this is to use the idea of virtual archives which is discussed below.

- Commercial content publishers (journal, software and music publishers, etc.) impose licence terms and conditions on the potential uses of their content. Even when content is freely available, it may still be covered e.g. by a copyleft, Creative Commons or similar licence. Increasingly, these conditions are implemented through Digital Rights Management software. Archives need to be aware of the impact of such DRM systems on their collection practices.

Opportunity: Support the increased use of technologies to ensure that content collected and held by archives satisfies the kinds of constraints mentioned above. Examine the possibility of using virtual archive technology to overcome some of the hurdles to selective accessibility. Digital Rights Management (DRM) tools must be taken into account, as they have the potential both to make the uses of content more clear, but also have the potential to restrict uses of content.

Challenges to Use

Many of the challenges mentioned above have, of course, a direct impact on the potential uses of web archives, in terms of the nature, scope and quality of the content made available to researchers. But in addition there are a number of specific challenges in the ways that researchers can go about making use of the available archive content, some of which present significant obstacles to wider adoption of web archives as a useful resource for research.

Visualisation and Navigation

Currently, most web archive interfaces are restricted to URL-based lookup, and simple searches based on metadata such as date ranges. Examples of such relatively limited interfaces include:

- The Wayback Machine[55] (developed by the Internet Archive) which provides only URL- and limited metadata-based search of the IA; it does not currently support text-based search (but see the discussion of NutchWAX below), nor does it have support for Google-style PageRank linkage analysis or for viewing collections are coherent, structured entities.

- WERA[56], supported by IIPC, an archive viewer application that gives an Wayback Machine-like access to small to medium-sized (up to 500 million documents) web archive collections, but also supports full text search and easy navigation between different versions of a web page, and a simple timeline view of archives.

- The Hanzo WARC tool suite[57], funded by the IIPC, a set of core libraries/APIs and command-line tools for full-text (based on Ferret[58]) and metadata-based search of archives in WARC format.

One fundamental problem is that of *scale*, both in terms of achieving adequate performance (without having the resources of the Google infrastructure) for browsing and searching. Another problem is that of finding the best metaphors and mechanisms to allow multi-scale (in the sense of Google Maps and Google Earth) overviews on complex collections, thereby allowing users to develop a sense of the context in which content was originally created. Interesting web visualisation tools which try to address this latter problem include:

- Touchgraph GoogleBrowser[59] (see Figure 3), which visualises the connectivity between websites and their relative sizes, as reported by Google's database of related sites, and the similar FacebookBrowser[60]

- Quintura[61] a tag cloud based visualisation and search tool

- walk2web[62] (see Figure 4) a graphical web visualisation tool which also supports social annotation features.

[55] http://www.archive.org/web/web.php
[56] http://archive-access.sourceforge.net/projects/wera/
[57] http://www.hanzoarchives.com/technology/open_source_projects
[58] http://ferret.davebalmain.com/
[59] http://www.touchgraph.com/TGGoogleBrowser.html
[60] http://www.touchgraph.com/TGFacebookBrowser.html
[61] http://www.quintura.com/
[62] http://www.walk2web.com/

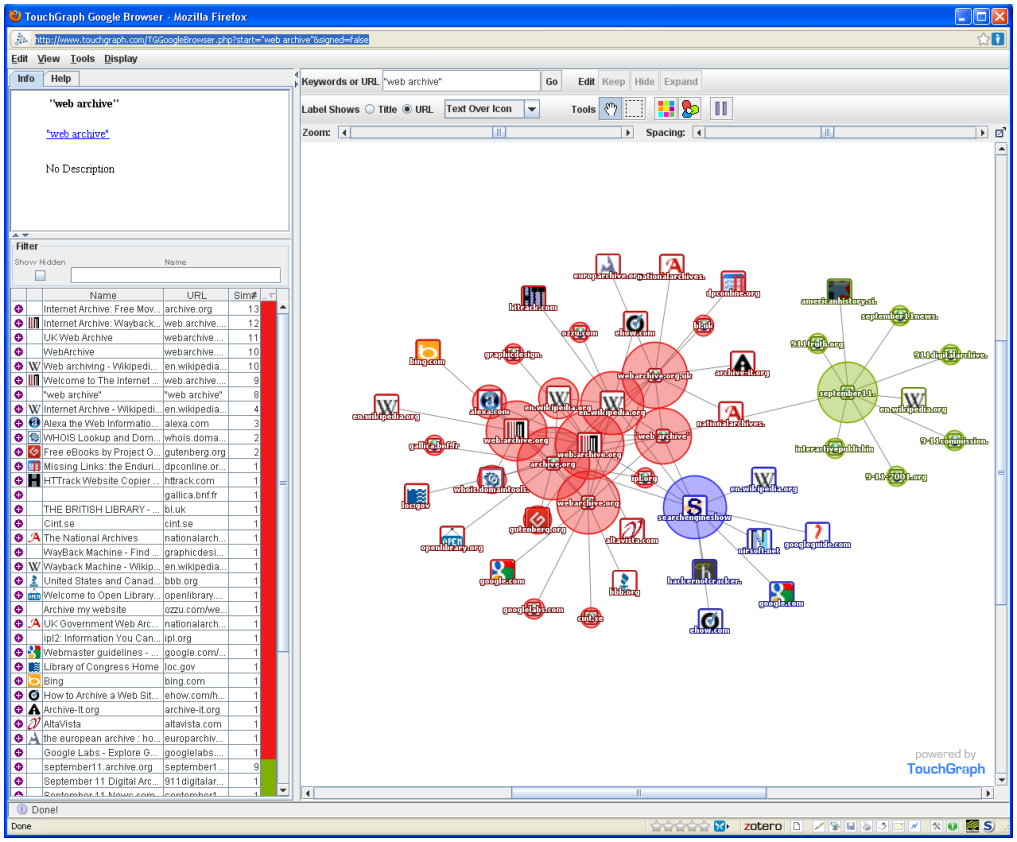

Figure 3: TouchGraph GoogleBrowser visualisation of the search term "web archive"

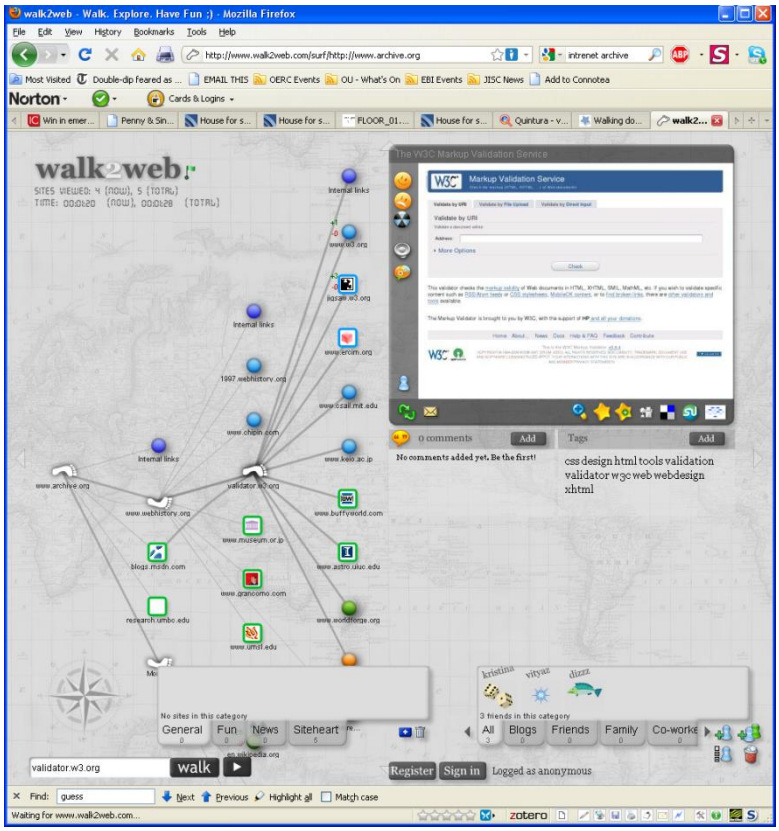

Figure 4: walk2web visualisation

Opportunity: Support development of tools like those described above which allow a much richer visualisation of web contexts, and allow complex, multi-scale and overlaid visualisations of web site content and relationships.

Search

As archives grow ever larger, browsing become less useful and sophisticated search functions become more important. Unfortunately, this is an area which, through constraints on resources, has received relatively little attention to date within the archiving community.

Examples of existing web archive search engines include:

- NutchWax[63], an open-source tool which extends the Nutch[64] web search engine for searching archives (currently only WARC/ARC files), supported by the IIPC and the IA. Its performance is relatively limited, but recent work to incorporate MapReduce/Hadoop-based algorithms and a Hadoop-style distributed file system into Nutch has improved performance considerably. One continuing major limitation of the Nutch search engine is its lack of support for Boolean operators or approximate string matching.

- Hanzo Search tools[65], part of the Hanzo WARCH tool kit, which provide an open-source toolkit (based on Ferret) for full-text search. Its performance also leaves something to be desired.

Few of these tools have the scalability required for growing archives: they are capable of managing and searching archives with a few hundred millions of documents, compared to Google which can scale by a factor of a 1000 or more beyond that.

Searching multimedia content is another significant challenge. While there are a variety of services (e.g. Google Images[66] and Google Goggles[67]) which claim to able to find content on the basis of image queries, such methods are not yet sufficiently precise to give accurate retrievals on large collections (Datta, Joshi, Li, & Wang, 2008). Extending still-image query-by-content to videos is an as-yet unsolved problem.

Opportunity: The archival search community seems to be fighting a losing battle in the face of the growing scale of available content. **The opportunity exists, especially given the available public Hadoop Cloud implementations, to move search to a much more ambitious level of performance at relatively modest cost.**

Semantic Web and Linked data

As described above, Web 3.0, Semantic Web and Linked Data concepts are rapidly becoming relevant to Web archive users. Linked Data can be viewed as content to be retrieved, or as metadata to be used to identify content to be retrieved. The Semantic Web community is putting considerable effort into Linked Data browsers, and examples include:

[63] http://archive-access.sourceforge.net/projects/nutch/index.html
[64] http://nutch.apache.org/
[65] http://code.google.com/p/search-tools/
[66] http://images.google.com/
[67] http://www.google.com/support/mobile/bin/answer.py?hl=en&answer=166331

- The DISCO Hyperdata Browser[68]
- OpenLink Data Explorer [69](a Firefox add-on)
- ObjectViewer[70]
- W3C Tabulator[71]
- Sindice[72]
- Marbles Linked Data Browser[73]
- Longwell Faceted RDF Browser[74]
- Gruff: AllegoGraph Triple Store Browser[75]

Opportunity: The tools being developed by the Semantic Web/Linked Data community could radically simplify the archival metadata management problem, and allow those metadata to be searched at scale and also used for data integration across collections in much more sophisticated ways.

Web Analytics

As archives become more ubiquitous and comprehensive, there is a growing need for tools to visualise and analyse the structure and evolution of web content across the entire web (see the discussion of web science below). Some initial attempts in this direction include:

- Hanzo has experimented with Guess, Graphviz and Hypergraph-based tools for graph visualisation. These were exploited in the JISC/NEH-funded OII/IA World Wide Web of Humanities project.[76]
- MediaCloud[77] is a comprehensive attempt to understand the temporal patterns of information percolation across the Internet. It provides tools to visualise the geographical and temporal appearance of information items

Opportunity: Support development of MediaCloud-style tools which allow users to query where ideas originated, and how they were propagated.

Multiple and Perspectives and Navigation: Temporal and Semantic Mashups

There is great interest in understanding the temporal evolution of collections ("how did it change from 2005-2009?" "what is different now?", "when was there most content on this topic"), but

[68] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/
[69] http://linkeddata.uriburner.com/ode/
[70] http://projects.semwebcentral.org/projects/objectviewer/
[71] http://dig.csail.mit.edu/2005/ajar/ajaw/About.html
[72] http://blog.sindice.com/2009/07/22/sigma-live-views-on-the-web-of-data/
[73] http://marbles.sourceforge.net/
[74] http://simile.mit.edu/wiki/Longwell
[75] http://www.franz.com/agraph/gruff/index.lhtml
[76] http://www.oii.ox.ac.uk/research/?id=48
[77] http://www.mediacloud.org/

doing this effectively has proved to be challenging. Google Replay[78] is an interesting example of timeline-based browsing. Yahoo! Pipes[79] provide tools for aggregating, manipulating and mashing-up content from web sites, and tools of this kind could be very useful for archive users.

Integrating content from different archives over different time periods requires, as described above, a much more powerful metadata model and tools for integrating and making inferences across multiple repositories. The Linked Data model is one such, supported by a huge body of research and tool-building within the Semantic Web community.

Opportunity: Support development of Linked Data-based metadata tools which allow constructing of complex mashups on the basis of temporal or rich semantic relationships. The archive community could also support the use Pipes to create mashups, and users could share Pipes modules for specific functions to support rapid building of customised processing pipelines.

Better emulation

The problem of how to render web objects in environments which emulate their original contexts (web browser version, fonts, codecs, etc.) is as yet unsolved (van der Hoeven, 2009). It is part of the larger and harder problem of preserving digital execution environments (including hardware and storage media).

KEEP:[80] the Keeping Emulation Environments Portable project is developing an Emulation Access Platform to enable accurate rendering of both static and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases and videogames.

Opportunity: There is a desperate need for emulation tools which will enable the continued viewing of content in obsolete formats, and using obsolete viewing platforms.

Better APIs and Web Services

Most of the emphasis in the archiving community has, understandably at this early stage, been on generic and monolithic tools of widespread appeal and use, but now is the time to pay attention to allowing users to build applications which address *their* specific needs. Very few if any of the tools built by the community use the web services paradigm, and therefore preclude the use of user-friendly workflow tools such as Taverna[81] to create complex analytical pipelines. Some semi-commercial tools, such as the Yahoo! Pipes tool mentioned above could be used as models. SOAP-based[82] or, better, RESTful[83] web service interfaces will allow rapid development of new applications.

Opportunity: Encourage the development of well-designed web service interfaces to allow rapid development of new applications for access to and analysis of archival content.

[78] http://googleblog.blogspot.com/2010/04/replay-it-google-search-across-twitter.html
[79] http://pipes.yahoo.com/pipes/
[80] http://www.keep-project.eu/ezpub2/index.php
[81] http://www.taverna.org.uk/
[82] http://en.wikipedia.org/wiki/SOAP
[83] http://en.wikipedia.org/wiki/RESTful

Stability of Citations

> *Authors increasingly cite web pages and other digital objects on the Internet, which can "disappear" overnight. In one study published in the journal Science, 13% of Internet references in scholarly articles were inactive after only 27 months. Another problem is that cited web pages may change, so that readers see something different than what the citing author saw. The problem of unstable web citations and the lack of routine digital preservation of cited digital objects has been referred to as an issue "calling for an immediate response" by publishers and authors.* (WebCite, 2010)

An interesting recent initiative to tackle this problem of unstable citations is to create "citation archives" (Davis, 2010) which create a permanent repository, build on persistent identifiers, which can be used by authors alongside, or instead of the original URLs, to refer to archives copies of cited documents. Examples of such repositories include the DACHS Citation Archive (Lecher, 2009) and WebCite.[84] WebCite is an on-demand archiving system for web references, which can be used by authors, editors, and publishers, to ensure that cited web material will remain available to readers in the future (WebCite, 2010). WebCite makes use of opaque URLs generated by the WebCite server when an object is submitted to the archive, and these URIs are guaranteed to be permanent and unique.

An alternative approach to WebCite's opaque URLs is to use persistent URLs,[85] which do not directly describe the location of the resource to be retrieved but instead describe an intermediate (more persistent) location (e.g. PURL.org[86]) which, when retrieved, results in redirection to the current location of the actual resource.

Opportunity: Make web archive contents look just like any other citable source. In this way, the boundary between current and historical materials can be blurred, and users can, as required, be redirected from live to archive sites (see below). Encourage the use of persistent URLs by web archives, to allow them to interoperate with Linked Data repositories (see below).

Creating Virtual Archives

Many recent studies have looked at the ways in which users go about creating thematic collections. The traditional way of doing this is directly by selection and collection using users' own resources, and whatever archival tools they feel comfortable with (Meyer, Carpenter, & Middleton, 2009).

An alternative approach would be one of creating views (in the sense of relational databases) into already-existing large scale archives. As discussed above, the drastic reduction in the costs of storage make feasible the idea of "collecting the haystack" and then using views to find and make accessible the "needles." Existing web content negotiation, transcoding, transclusion and mashup technologies provide the first stepping-stones to implementation of this concept.

Opportunity: Promote the development of tools for creating views (selected subsets of contents) into large-scale archives.

[84] http://www.Webcitation.org/
[85] http://en.wikipedia.org/wiki/Persistent_Uniform_Resource_Locator
[86] http://www.purl.org/docs/index.html

Blurring the Distinction between Live and Archive

Increasingly, search engines such as Google support caching of web content (usually limited to a few months of retrospective, primarily as a means to protect against content loss), and this trend is likely to accelerate (leading to the possibility that Google and its competitors will themselves eventually be major players in web archiving).

Users of web archives should be able to move transparently between current, live content and historical content, and all search and visualisation functions should work indifferently. This is a long way from being achieved, however, given the mindset of the archiving community and the limited resources available to make this possible. Short of waiting for Google, Bing and others to implement this functionality at their own pace and on their own scope, the web archive community should be active in promoting this blurring of distinctions.

Opportunity: The same technologies used for creating virtual archives could be used to blur the distinction between live and historical content.

Sharing and Interoperability

As more institutions undertake archiving, there is an increasing need for tools which support browsing and search across multiple archives. The Archives Hub[87] (Stevenson & Ruddock, 2010) is a JISC-funded service enabling users to search across 200 repositories. It uses the Library of Congress Encoded Archival Description (EAD),[88] a machine-readable standard for inventories, registries and indexes

This activity emphasises the need for standards in resource discovery, indexing, access methods and document formats. The formation of the UK Archives Discovery network (UKAD)[89] is an encouraging first step to addressing these needs. UKAD is working to promote the capacity for a cross-searching capability across the UK archive networks and online repository catalogues, to support resource discovery through the promotion of relevant national and international standards and to support the development and use of name authorities.

As with all data interoperability efforts, there are major challenges in achieving agreement on naming and index terms. While formal ontologies have been built for a wide variety of domains, especially in the sciences, the development of such resources is highly labour-intensive and time-consuming. There are some promising developments in using text mining methods to populate ontologies (e.g., Witte, Khamis, & Rilling, 2010) which suggest that this process can to some extent be automated. Folksonomies developed with the participation of user communities also offer promising results in some areas.

The Semantic Web community has long recognised the importance of persistent identifiers, and a variety of services such as PURL.org have developed to support their use, but the development of consistent naming schemes or URI patterns still requires considerable effort and coordination (UK Chief Technology Officer Council, 2009).

[87] http://archiveshub.ac.uk/
[88] http://www.loc.gov/ead/index.html
[89] http://archivesnetwork.ning.com/main/authorization/signIn?target=http%3A%2F%2Farchivesnetwork.ning.com%2F

Opportunity: Continue JISC support for the Archives Hub, but encourage investigation of Linked Data standards for building and linking metadata repositories. Emphasis should be places on coordinating with other efforts to develop consistent URI naming schemes for public sector domains.

Users as Creators: Social Annotation

The whole concept of Web 2.0 blurs the distinction between content creators and content consumers. Blogs now universally allow readers to comment and (for better or worse) many mainstream media web sites now do the same. The social media universe now includes a wide variety of public and commercial social annotation (comment, tagging, social bookmarking, rating and recommending, etc.), tools (such as Zotero, Connotea, del.icio.us, Digg, Reddit, etc. as well as wikis, of course) and these are widely used.

Opportunity: Users of web archives should be able to use these kinds of tools to annotate and mark archive contents in ways which allow sharing with other readers. Currently, no major web archives support this functionality.

Challenges to Web Science

Understanding Development of Web Content

Even though the web is now 20 years old, our understanding of the dynamics and life cycle of web content is still relatively under-developed. This is especially true as we move into the area of Web 2.0 and its emphasis on user-created content, collaboration and sharing.

Wikis provide powerful and widely-used platforms for sharing knowledge, and for community curation of knowledge bases. But at present, we have only a very limited understanding of the sociology of such collaborative ventures, and are unable to answer simple questions such as “who is the most influential contributor on this topic?” or “who should get credit for this thread of ideas?” While the changelogs contain basic information about what edits were made, when and by whom, there are more sophisticated inferences which could be drawn from them which could begin to answer theses kinds of questions.

Opportunity: Support research into the rich vein of information that is made available by collaborative web sites concerning the dynamics of information creation and sharing. Support development of techniques and tools for in-depth analysis of how community curation actually works, and what could be done to make it better.

Understanding how Information Propagates

The web, of course, is not a single information space, but rather a complex and inter-related family of such spaces (commercial and personal web sites, mainstream media sites, the blogosphere, Twitter, etc.), each with its own points of view and emphases, and each with its own dynamic.

To date there have been relatively few wide-raging studies on how information propagates within and between these spaces. It is not known, for example, how trends or memes which have their origins in the blogosphere percolate into the mainstream media (Leskovec, Backstrom, & Kleinberg,

2009; Lohr, 2009). Some simple tools such as the Google Zeitgeist[90] and Twitter's Trending Topics[91] list serve to give snapshots of those topics which are getting most attention at any instant; tools such as Google Replay[92] allow some timeline visualisations of the popularity of topics. More interesting approaches such as MediaCloud[93] (Cohen, 2009) represent attempts to understand the temporal patterns of information percolation across the Internet. They provide tools to visualise the geographical and temporal appearance of information items. Commercial tools, such as Recorded Future's Temporal Analysis Engine[94] also provide useful visualisations (see Figure 5) of the ebb and flow of content across a wide range of web content.

Figure 5: Recorded Future Visualisation of News Timeline

Opportunity: Support approaches, such as those of MediaCloud and Recorded Future, to understand the dynamics of spatio-temporal percolation of ideas in the web.

Understanding the topology, dynamics and evolution of the Web Graph

In recent years, the study of the web as an information artefact in its own right has come to occupy an important place in computer science research (Hendler, et al., 2008). While it was recognised early on that studying the properties of the web graph (to decide, for example, whether it represents a scale-free network) could give great insights into the structure and evolution of the web, such studies are still not able to deal with seemingly crucial questions such as:

- the influence of user-dependent state (e.g. on how a resource is presented to a user in the presence of a particular cookie, or via a particular RESTful URL)
- how to represent other forms of dynamic and deep web content

[90] http://www.google.com/intl/en/press/zeitgeist/index.html
[91] http://twitter.com/trendingtopics
[92] http://googleblog.blogspot.com/2010/04/replay-it-google-search-across-twitter.html
[93] http://www.mediacloud.org/
[94] https://www.recordedfuture.com/

- how to represent the micro-behaviour of user interactions, especially those which are executed locally within browsers or apps
- how to deal with dynamically varying constraints on access (e.g. to content only accessible to friends in Facebook)
- whether subsets of the web graph (e.g. those in collaborative sub-networks such as wikis) have different properties from the web as a whole.

Opportunity: Support development of tools and methodologies for using web archives to study the historical dynamics and evolution of the web graph.

The Web as a Social Machine

Especially with the advent of user-created content, it may be useful to think of the web not as a distributed information repository, but rather as a "social machine" (Berners-Lee & Fischetti, 1999; Hendler & Berners-Lee, 2010) which supports the social and professional interactions of communities, whether to create new knowledge or to solve problems collectively (especially problems which involve interaction between the online and offline worlds). This view stresses the interaction between technology and sociology in ways which promote emergent behaviour from large communities, and the influences of factors such as trust and reputation on these behaviours.

Opportunity: Work on social machines is just beginning, and there is a valuable opportunity to ensure that web archiving captures the kinds of information which makes the study of such artefacts possible.

Conclusion and Summary of Opportunities

The discussion above makes clear that there is a wide range of challenges facing the web archive community and its users. But these challenges also provide opportunities for funding agencies to promote a much more aggressive and ambitious programme of collection and use of web (and more generally, Internet) content. The rapidly-changing nature of the web sharpens up many of the issues which have, until now, received relatively scant attention by the web archiving community. To be fair, this lack of attention seems to be due principally to the ludicrously modest resources which have been available to the community over the last two decades (certainly by comparison with the resources used to create the content in the first place, and the resources brought to bear on web search by Google, Bing, Yahoo! and others).

In our view, the principal opportunities which should be actively pursued by the funding agencies can be summarised as below.

Collection

1. The creation of a much better set of tools that would allow archivists to select, collect and validate content easily; these tools should provide:
    a. better visualisations of single web sites (e.g. using Sitemaps) and collections of web sites.
    b. more flexible ways (APIs, web service interfaces) of constructing workflows (using e.g. Taverna) or specific collection applications
    c. better tools for capturing, analysing and searching metadata (which would lead to better ways of dealing with the complex structures of collections); these should support the use of the Linked Data model as a more flexible paradigm
    d. significantly new functionality for dealing with multi-media content.
2. Continue to support collaborative efforts. The IIPC has been instrumental in the creation of tools and standards that have been widely adopted in web archiving communities, and such consortium efforts remain an important means of creating collections, tools, and services that are useful to a wide range of researchers. But such efforts have a time-course of their own which is not well-matched to the rate of evolution of the web.
3. Recognise that, given the rapidly-decreasing cost of storage, effort may be better spent on "collecting the haystacks" rather than putting huge efforts into identifying the "needles" a priori; this would have the highly-desirable side-effect of allowing much easier creation of thematic archives by individual users, making use of concepts such as *views* and *virtual archives*.
4. Promote cloud storage as a cost-effective, flexible way to handle growing data volumes.
5. Support more sophisticated use of technologies (which may include Digital Rights Management) to address the growing complexity of copyright and privacy issues which face both archivists and users.
6. Better policy coordination to reduce overlap between collections, and the use of e.g. the ArchiveHub portal to publish holdings.

Dealing with New and more Complex Types of Content

1. Recognise that the deep web and dark webs pose significant new challenges to archiving, especially in capturing interactive behaviours.
2. Tackle the hard problem of state and executable content, which can cause the appearance of content to change dynamically.

3. Recognise the growing importance of non-PC-centric views of the web, especially of dynamically-reconfigured content served to mobile devices.
4. Support development of tools and techniques to capture Web 2.0 content, recognising the enormous volumes of data involved, and the wide variety of data representations used by different sites.
5. Recognise the growing importance of data (as represented e.g. by Web 3.0 and especially Linked Data); find ways to collect and search these rapidly-growing universes.

More Powerful and Flexible Access to Archives

1. Provide richer APIs and web services interfaces to archives, to allow users to develop their own query interfaces, visualisations and mashups; support the idea of virtual archive technology to allow users to create their own thematic collections from generalised archives.
2. Recognise that current archive search tools are not sufficiently scalable or flexible to deal with growing volumes and new kinds of content.
3. Support the development of much more powerful visualisation tools that allow navigation through large and complex archives, and support temporal and semantic mashups within and across archives.
4. Given increased use of Linked Data models for metadata, support the application of existing Linked Data browsing and data integration tools to archival content.
5. Recognise that the distinction between archival and live web content is an artificial one, and that tools should transparently support navigation, search and analysis across content irrespective of its nature.

Supporting Web Science

1. Support the development of much richer tools for studying the structure and evolution of the web graph, recognising that simple-minded approaches fail to capture much of the social and behavioural contexts in which content is created and used.
2. Promote efforts to understand, and improve, community curation of web content, and the ways in which credit can be assigned, and impact measured, in such collaborative content.
3. Support the development of tools to analyse how information propagates between different web domains
4. Recognise the growing importance of the web as a social machine at the interface between the world of information, the world of human cooperation and the physical world (the “Web of Things”), and develop means to capture and analyse these heretofore neglected aspects of the place of the web in human collective behaviour.

To sum up, the current state of web archiving is not good, and is falling ever further behind the rapid evolution of its object of collection. Archivists face huge challenges in collecting rapidly-evolving and ever-growing volumes of content, while users struggle to comprehend what content is available and how to make use of it. If we are not to lose forever the opportunity to capture one of the most significant recent developments in human cultural history, major new investment, and a much more ambitious programme of development work is needed.

Appendix A: Interviews

For this project, we supplemented desk research with interviews with 17 stakeholders in the web archiving community. We are grateful to the following individuals for generously helping us to better understand how archivists and researchers are engaging with web archives.

Niels Brügger
Associate Professor, Department of Information and Media Studies
Aarhus University, Denmark

Richard Davis
Repository Service Manager
University of London Computer Centre, United Kingdom

Katrien Depuydt
Head of the Language Database Department
Institute for Dutch Lexicology, The Netherlands

Kirsten Foot
Associate Professor of Communication
University of Washington, United States of America

Wendy Gogel
WAX Project Manger
Harvard University Library, United States of America

Alison Hill
Curator, Web Archiving, Modern British Collections
The British Library, United Kingdom

Helen Hockx-Yu
Web Archiving Programme Manager
The British Library, United Kingdom

Hanno Lecher
Librarian, China Studies
Leiden University, The Netherlands

Julien Masanès
Director
European Archive, France

Frank McCown
Assistant Professor of Computer Science
Harding University, United Kingdom

Mark Middleton
CEO, Hanzo Archives, United Kingdom

Martin Moyle
Digital Curation Manager
University College London (UCL) Library Services, United Kingdom

Kris Carpenter Negulescu
Director of the Web Archive
Internet Archive, United States of America

Ed Pinsent
Digital Archivist/Project Manager
University of London Computer Centre, United Kingdom

Steve Schneider
Professor & Interim Dean, School of Arts & Sciences
SUNY Institute of Technology, United States of America

René Voorburg
Crawl-engineer & Coordinator of web archiving
Acquisition and Processing Division – E-depot
Koninklijke Bibliotheek, The National Library of the Netherlands, The Netherlands

Max Wilkinson
Datasets Programme Technical Lead
British Library, United Kingdom

References Cited

Allinson, J. (2006). *OAIS as a Reference Model for Repositories: An Evaluation*. Bath: UKOLN, University of Bath. Retrieved from http://www.ukoln.ac.uk/repositories/publications/oais-evaluation-200607/

Anderson, C., & Wolff, M. (2010, September). The Web is Dead. Long Live the Internet. *Wired*.

Antonescu, M.-D., Guttenbrunner, M., & Rauber, A. (2009). Documenting a Virtual World: A Case Study in Preserving Scenes from Second Life. In *Proceedings of the The 9th International Web Archiving Workshop (IWAW 2009)*, Corfu, Greece. Retrieved from http://www.iwaw.net/09/IWAW2009.pdf

Barbosa, L., & Freire, J. (2007). An Adaptive Crawler for Locating Hidden-Web Entry Points. In *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, Banff, Alberta, Canada. Retrieved from http://portal.acm.org/citation.cfm?id=1242632&dl=GUIDE

Benczúr, A., Siklósi, D., Szabó, J., Bíró, I., Fekete, Z., Kurucz, M., et al. (2008). Web Spam: A Survey with Vision for the Archivist. In *Proceedings of the 8th International Web Archiving Workshop (IWAW 2008)*, Aarhus, Denmark. Retrieved from http://iwaw.net/08/IWAW2008-Benczur.pdf

Bergman, M. K. (2001). The deep web: Surfacing hidden value. *Journal of Electronic Publishing, 7*(1).

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web*. New York: Harper Collins.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(34-43).

CCSDS. (2002). *Consultative Committee for Space Data Systems Reference Model for an Open Archival Information System (OAIS), CCSDS Blue Book 650.0-B-1*. Reston, VA: CCSDS. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

Cohen, P. (2009, 4 August). Hot Story to Has-Been: Tracking News via Cyberspace, *New York Times,* p. C1. Retrieved from http://www.nytimes.com/2009/08/05/arts/05cloud.html

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image Retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys, 40*(2), 1-60.

Davis, R. (2010). Moving Targets: Web Preservation and Reference Management. *Ariadne*(62).

Dougherty, M., Meyer, E. T., Madsen, C., Van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art*. London: JISC.

Guenther, R., & Myrick, L. (2006). Archiving Web Sites for Preservation and Access: MODS, METS and MINERVA. *Journal of Archival Organization, 4*(1-2), 141-166.

Guha, R. V., Kumar, R., Raghavan, P., & Tomkins, A. (2004). Propagation of trust and distrust. In *Proceedings of the 13th International Conference on World Wide Web (WWW 2004)*, New York. Retrieved from http://portal.acm.org/citation.cfm?id=988672.988727&coll=portal&dl=ACM&type=series&idx=SERIES968&part=series&WantType=Proceedings&title=WWW

Guthrie, D. (2008). *Unsupervised Detection of Anomalies in Text.* Ph.D. thesis, University of Sheffield, Sheffield.

Guthrie, D., Guthrie, L., Allison, B., & Wilks, Y. (2007). Unsupervised anomaly detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Retrieved from http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-262.pdf

Hendler, J., & Berners-Lee, T. (2010). From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence, 174*(2), 156-161.

Hendler, J., Shadbolt, N., Berners-Lee, T., & Weitzner, D. (2008). Web Science: An Interdisciplinary Approach to Understanding the Web. *Communications of the ACM, 51*(7), 60-69.

Herman, I. (2010). W3C Semantic Web Activity. Web page retrieved 01 August, 2010, from http://www.w3.org/2001/sw/

Hunter, J., Cole, T., Sanderson, R., & Ven de Sompel, H. (2010). The Open Annotation Collaboration: A Data Model to Support Sharing and Interoperability of Scholarly Annotations. In

Proceedings of the Digital Humanities 2010 Conference, King's College, London. Retrieved from http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-860.pdf

Lavoie, B., & Gartner, R. (2005). *Preservation Metadata (Report 05-01)*. York, UK: Digital Preservation Coalition. Retrieved from http://www.dpconline.org/component/docman/doc_download/88-preservation-metadata

Lecher, H. (2009). Web Archive and Citation Repository in One: DACHS. In *Proceedings of the Workshop on Missing Links: The Enduring Web*, London. Retrieved from http://www.dpconline.org/component/docman/doc_download/395-0907lechermissinglinks

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, Paris, France. Retrieved from http://www.cs.cornell.edu/home/kleinber/kdd09-quotes.pdf

Linden, T. (2010). 2009 End of Year Second Life Economy Wrap up (including Q4 Economy in Detail). *Second Life Blog*. Web page retrieved 28 July, 2010, from http://blogs.secondlife.com/community/features/blog/2010/01/19/2009-end-of-year-second-life-economy-wrap-up-including-q4-economy-in-detail

Lohr, S. (2009, 12 July). Study Measures the Chatter of the News Cycle, *The New York Times,* p. B1. Retrieved from http://www.nytimes.com/2009/07/13/technology/internet/13influence.html

McCown, F., & Nelson, M. L. (2009). What happens when facebook is gone? In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, Austin, TX. Retrieved from http://portal.acm.org/citation.cfm?id=1555400.1555440

Meyer, E. T., Carpenter, K., & Middleton, M. (2009). *World Wide Web of Humanities: Final Report to JISC*. London: JISC. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/digitisation/humanitiesfinalreport.pdf

Morrison, H. (2007). Rethinking collections - Libraries and librarians in an open age: A theoretical view. *First Monday, 12*(10).

Nelson, M. L., Smith, J. A., Del Campo, I. G., Van de Sompel, H., & Liu, X. (2006). Efficient, automatic Web resource harvesting. In *Proceedings of the 8th ACM International Workshop on Web Information and Data Management (WIDM 2006)*, Arlington, VA. Retrieved from http://public.lanl.gov/herbertv/papers/f140-nelson.pdf

PVW. (2010). Preserving Virtual Worlds: Meta Data Schema Development. Web page retrieved 2010, 28 July, from http://pvw.illinois.edu/pvw/?page_id=25

Ras, M., & van Bussel, S. (2007). Web Archiving User Survey. Retrieved from http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/KB_UserSurvey_Webarchive_EN.pdf

Riley, J. (2010). Glossary of Metadata Standards. Retrieved from http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards_glossary_pamphlet.pdf

Risse, T. (2009). From Web Page to Living Web Archive. In *Proceedings of the Workshop on Missing Links: The Enduring Web*, London. Retrieved from http://www.dpconline.org/component/docman/doc_download/403-0907rissemissinglinks

Spaniol, M., Mazeika, A., Denev, D., & Weikum, G. (2009). 'Catch me if you can': Visual Analysis of Coherence Defects in Web Archiving. In *Proceedings of the 9th International Web Archiving Workshop (IWAW 2009)*, Corfu, Greece. Retrieved from http://www.iwaw.net/09/IWAW2009.pdf

Stevenson, J., & Ruddock, B. (2010). Moving towards Interoperability: Experiences of the Archives Hub. *Ariadne, 63*(April).

Tofel, B., & Vahlis, E. (2006, 14 August 2006). Leverage browsers for link-extraction: Browser Monkeys. Web page retrieved 29 July, 2010, from https://webarchive.jira.com/wiki/display/SOC06/Leverage+browsers+for+link-extraction

UK Chief Technology Officer Council. (2009). *Designing URI Sets for the UK Public Sector: A report from the Public Sector Information Domain of the CTO Council's Cross-Government Enterprise Architecture (Interim paper, Version 1.0)*. London: Chief Technology Officer Council. Retrieved from www.cabinetoffice.gov.uk/media/301253/puiblic_sector_uri.pdf

van der Hoeven, J. (2009). Emulating Access to the Web 1.0. In *Proceedings of the Workshop on Missing Links: The Enduring Web*, London. Retrieved from http://www.dpconline.org/component/docman/doc_download/400-0907vanderhoevenmissinglinks

WebCite. (2010). WebCite web page. Web page retrieved 28 July, 2010, from http://www.Webcitation.org/

Witte, R., Khamis, N., & Rilling, J. (2010). Flexible Ontology Population from Text: The OwlExporter. In *Proceedings of the International Conference on Language Resources and Evaluation*, Malta. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/932_Paper.pdf

Wright, A. (2009). Exploring a 'Deep Web' That Google Can't Grasp, *The New York Times,* p. B4. Retrieved from http://www.nytimes.com/2009/02/23/technology/internet/23search.html