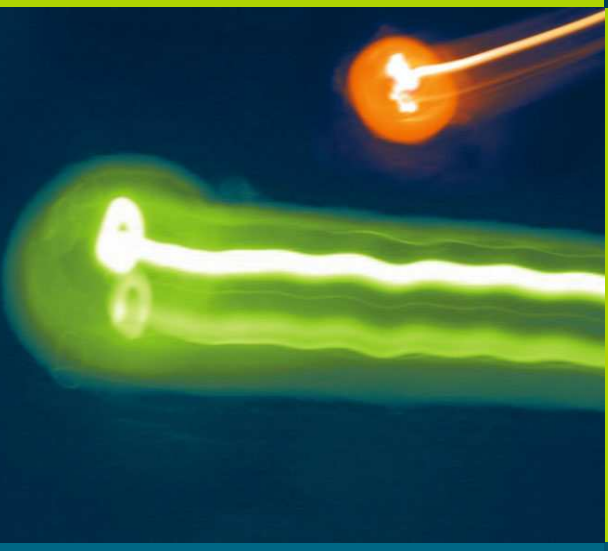


Julien Masanès (Ed.)

Web Archiving

Springer

Web Archiving

Julien Masanès

Web Archiving

With 28 Figures and 6 Tables

Springer

Author
Julien Masanès
European Web Archive
25 rue des envierges
75020 Paris, France
julien.masanes@bnf.fr

ACM Computing Classification (1998): H.3, H.4, I.7, K.4
Library of Congress Control Number: 2006930407

ISBN-10 3-540-23338-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-23338-1 Springer Berlin Heidelberg New York

Typesetting by the author and SPi

Cover design: KünkelLopka, Heidelberg

Printed on acid-free paper SPIN: 11307549/45/2145/SPi 5 4 3 2 1 0

Contents

1 Web Archiving: Issues and Methods

Julien Masanès

European Web Archive
julien@iwaw.net

1.1 Introduction

Cultural artifacts of the past have always had an important role in the formation of consciousness and self-understanding of a society and the construction of its future. The World Wide Web, Web in short, is a pervasive and ephemeral media where modern culture in a large sense finds a natural form of expression. Publications, debate, creation, work, and social interaction in a large sense: many aspects of society are happening or reflected on the Internet in general and the Web in particular.[1] Web preservation is for this reason a cultural and historical necessity. But the Web is also different from the previous publication systems to necessitate a radical revision of traditional preservation practices.

This chapter presents a review of issues that Web preservation raises and of methods that have been developed to date to overcome them. We first discuss arguments against the necessity and possibility of Web archiving. We then try to present the most salient differences that the Web presents from other cultural artifacts and draw their implications for preservation. This encompasses the web's cardinality, the Web considered as an active publishing system, and the Web considered as a hypermedia collectively edited or a global cultural artifact. For each of this aspect of the Web, we discuss preservation possibilities and limits. We then present the main methodological approaches for acquisition organization and storage of Web content. Chapters 2, 4, and 5 provide further details on methodologies and tools for acquisition of content, and Chaps. 6–8 focus on access, mining, and preservation of Web content. The two final chapters of this book present case studies: the Internet Archive, the largest Web archive in the world (Chap. 9) and DACHS a research-driven selective Web archive (Chap. 10). This chapter can thus be considered as a general introduction to the book. Finally, it provides a presentation of initiatives in this domain and proposes a taxonomy of Web archives to map the current state of Web preservation.

[1] On the social dimension of networks and a discussion of the far reaching consequences that it entails, see Castells, (1996), Levy (1997), Hine (2000).

1.2 Heritage, Society, and the Web

1.2.1 Heritage Preservation

The concept of collective heritage, encompassing every possible human artifact from architectural monuments to books, is relatively new and can be dated from the twentieth century albeit related preservation activities (as systematically and voluntary organized ones) appeared earlier. Form, goals, and efficiency of heritage preservation have varied significantly with time and medium considered and it is not the ambition of this chapter to summarize this rich evolution. Let us just recall that from religious intellectual preparation (with the Vivarium library of Cassiodorus, Riché 1996) to collection building as a sign of power (see invention of modern museum by the Medicis in Florence late fifteenth century) to systematic state-control and national culture preservation (see invention of legal deposit Francois 1er), various motivations drove to systematic collection and preservation of cultural artifacts in history.

In modern time, archives in general tend to be more and more inclusive (Osborn 1999). As Mike Featherstone explains:

> Archive reason is a kind of reason which is concerned with detail, it constantly directs us away from the big generalization, down into the particularity and singularity of the event. Increasingly the focus has shifted from archiving the lives of the good and the great down to the detail of mundane everyday life. (Featherstone 2000).

In fact, the facility that Web brings for publishing, offers a unique source of this type of content that modern archive reason tend to praise. We could therefore assume that legitimacy for Web archiving is well established and acknowledged. Despite this, preserving the Web has been questioned and is not yet accepted by all. Argument against web archiving can be classified in three categories: those based on the quality of content found on the web, the ones that consider the Web is self-preserving and the ones that assume archiving the Web is not possible.

1.2.1.1 Not Good Enough?

The first category comprises arguments on Web content quality allegedly supposed to not meet required standards for preservation. This position has long been held by some professionals of the printing world (publishers, librarians) and went along with a larger sense of threat posed by this new media to their existence in general. It is usually associated with concerns about the vast amount of information the Web represents and a lack of

knowledge about Web archiving methods and costs. Advocate of this position are aware of the migration of the publication system online, and they which to continue preserving the publishing industry's output online. But they refuse to expand the boundaries of what is preserved as much as the Web has expanded the limits of what is "published". The economic equation of physical production of carrier for knowledge (book serials, etc.) inherited from the Gutenberg revolution, should, according to their view, continue to set limits to what should be preserved, even at a time where this equation is deeply modified. Historically, the fact that what could be published was limited by physical costs (including production but also transport, storage and handling costs) gave birth to the necessity for filtering, what the publishing system has accomplished for more than five centuries. But this is not any longer the case, and the relatively stable equilibrium inherited from the fifteenth century is broken. The development of the Web has dramatically increased the volume of what can be published as well as the number of potential "publishers" or content creators by dropping publications costs to almost nothing. The discussion on quality appraisal, inevitably subjective, is actually hiding the real debate about the expansion of the publishing sphere.

Although the growth of serial publication at the end of the nineteenth century is not comparable in size to the current revolution, it shares some characteristic (emergence of a new type of publication with a new temporality and a questioned intellectual status) and raised the same reactions. It took some times to the library community for instance to accept this type of publication in their shelves as well as in their heart. As Fayet-Scribe (2000) has shown for the case of France, the specific descriptive treatment that it required at the article level was, for this reason, neglected by this community and gave rise to an entire new sector of information management beside libraries (documentation, scientific literature indexing). The debate on archiving the Web shares some similarities with this episode. It remains to be seen if it will end in the same manner.

The filtering function, although not any longer required to allocate efficiently resources of physical production of carrier for knowledge, is, however, not entirely disappearing. It is rather shifting from a central role to a peripheral one, still needed in some spheres (for instance academic validation) and experiencing new forms (ex Wikipedia, slashdot, impact bogosphere).

As Axel Bruns explains:

> The repercussions of the emergence of this interactive and highly participatory mass medium continue to be felt. If everyone is, or at least has the potential to be, a publisher, what are the effects on existing publishing institutions? If information available on the Web can be

> easily linked together in a wide variety of combinations, what is the effect on traditional publishing formats? If there is a potential for audiences on the Web to participate and engage interactively in the production and evaluation of content, what happens to established producer and consumer roles in the mass media? (Bruns 2005)

With regards to preservation, this has also to be considered seriously. One thing is sure: it is a utopia to hope that a small number of librarians will replace the publisher's filter at the scale of the global Web. Even if they have a long tradition in selecting content, they have done this in a much more structured environment that was also several orders of magnitude smaller in size. Although this is still possible and useful for well-defined communities and limited goals (see Chap. 3 on selection methodologies and Chap. 10 on DACHS, a research-driven Web archive, see also Brügger (2005)), applying this as a global mechanism for Web archiving is not realistic. But the fact that manual selection of content does not scales to the Web size is not a reason for rejecting Web Archiving in general. It is just a good reason to reconsider the issue of selection and quality in this environment.

Could it be based on a collective and highly distributed quality assessment? Such an assessment is implicitly made at two levels: Web users by accessing content, creators by linking content form their pages (we do not consider here the judgment made by creator themselves before putting content online, that if used as a selection criteria, would mean just archiving everything). It could also be made explicitly by the multiplication of active selectors.

Let us consider users access first. The expansion of the online publication's sphere beyond what the economic capacity allowed for physical printing has other consequence: the mechanical drop in average number of readers of each unit of published content. Some pages are even not read by any human nor indexed by any robot at all. Boufkhad and Viennot (2003) have shown using the logs and file server of a large academic website that 5% of pages were only accessed by robots, and 25% of them were never accessed at all. This means that distribution of access to online content exhibits a very long tail.

But this evolution is not entirely new in modern publishing. The growth and high degree of specialization of serial publications already shows the same pattern of access. Is this an argument for not preserving serials? At least in most countries, legal deposit systems do preserve publication independently of how much they are being used. This provisions the indeterminacy of future reader's interests.

It is certainly possible for preservationists to evaluate usefulness (as measured by access) of online content for the present as well as trying to foresee it for the future as long as it is done for well-defined user communities. Access patterns can also be used for driving global archiving systems: it is the case of the main Web archive so far, the collection of the Internet Archive donated by Alexa, which use access patterns to determine depth of crawl for each site (see Chap. 9, Kimpton et al. (2006)). It can also be driven by queries sent to search engine (Pandey and Olston 2005). But the key question for Web archives would then be: how to get this information, and which threshold to use? Traffic information is not publicly available and search engines, following Alexa's innovation, get it from the millions of toolbars installed in browsers that pass user's navigation information to them. Where could archiving institutions get it as they do not offer search functionalities themselves? What should the threshold be? Should it be applied at the page or the site level (Alexa use it at the site level)? Would it constrain depth of crawl only (which means that at least the first level of each site will be captured in all cases)? Even if this criteria raises lots of practical implementation issues, it has the merit of taking as driver for archiving focus, the input of millions of users and not small committees, which is well adapted to the Web publication model itself.

The other criterion is the level of importance as measured by the inlinking degree of a page (or a site). It has been argued (Masanès 2002) that this is a pertinent equivalent in a hypertext environment of the degree of publicity that characterizes traditional publication and it has the advantage of being practically usable by mining the linking matrix of the Web (Page et al. 1998; Abiteboul et al. 2002, 2003; Pastor-Satorras and Vespignani 2004). It is another way of aggregating the quality assessment made, not by users, but by page (and links) creators. This distributed quality appraisal model is both well adapted to the distributed nature of publication on the Internet and practically possible to implement.

Finally, it is also possible to scale up by involving more and more participants in the task of selecting material to be archived. This can be done by involving more institutions in the task and facilitating this by providing archiving services that handle the technical part of the problem. This is proposed by the Archive-it service of Internet Archive that was launched in 2006. It enables easy collection set-up and management for libraries and archives that can't invest in the operational infrastructure needed for Web archiving.

Another possible evolution is the generalization of this to enable every Web user to participate actively if she or he wants, in the task of archiving the Web. The main incentive for users is, in this case, to organize their own personal Web memory to be able to refer back later to stable content,

but also to mine it and organize it as way to fight the "lost in cyberspace" syndrome. Several user studies actually show that keeping trace of content visited is essential to many users (Teevan 2004), but also that they use inefficient methods for this (Jones et al. 2001, 2003). Personal Web archive, recording user's trace on the Web could enable a personal and time-centric organization of Web memory (Rekimoto 1999; Dumais et al. 2003; Ringel et al. 2003).

Several online services (Furl, MyYahoo) already proposed personal Web archiving at the page level, combined with tagging functionalities. Hanzo Archives service allows extended scoping (context, entire site) as well as mashing up archiving functionalities with other tools and services (blogs, browsers, etc.) through an open API. It will be extended further with an archiving client with P2P functionalities that will dramatically extend possibilities for users to record their Web experience as part of their digital life (Freeman and Gelernter 1996; Gemmell et al. 2002) On potential use of user's cache in a Peer to Peer Web archive see also (Mantratzis and Orgun 2004).

It remain to be seen if this extension and democratization of the archiving role can expand like commentary and organization of information has been with the development of tagging (Golder and Huberman 2005) and blogging systems (Halavais 2004; Bruns 2005). But if it does, there could be a valuable help and input for preservation institutions, that can take long-term stewardship of this content.

As we have seen, arguments against Web archiving based on quality are grounded on the assumptions that 1/quality of content is not sufficient beyond the sphere of traditionally edited content, and that 2/only manual, one-by-one selection made by preservationists could replace the absence of publisher's filtering (approach that just cannot scale to the size of the Web, as all would agree Phillips (2005)). These two arguments shows lack of understanding of the distributed nature of the Web and how it can be leveraged to organize its memory at large scale.

1.2.1.2 A Self-Preserving Medium?

The second category of arguments holds that the Web is a self-preserving medium. In this view, resources deserving to be preserved will be maintained on servers, others will disappear at the original creator's will. As the first type of argument on quality was mostly found in the library world, this one finds most of its proponents in the computer science world. Although it was strongly supported in the early days, we have to say that, as time goes and content disappears from the Web, it is less the case. Many studies document the very ephemeral nature of Web resources defeating

the assertion that the Web is a self-preserving medium (see for instance Koehler (2004) and Spinellis (2003) for recent review of the literature on the subject). Studies show that the average half-life of a Web page (period during which half of the pages will disappear) is only two years. These studies focus on availability of resources at the same URL, not potential change they can undergo. Some also did verify the content and measured the rate of change. Cho and Garcia-Molina (2000) found a half life of 50 days for average Web pages, (Fetterly et al. 2003) showed how this rate of change is related to the size and location of the content.

They are many reasons why resources tend to disappear from the Web. First, it is the time limitation of domain name renting (usually 1–3 years) that puts, by design, each Web space in a moving and precarious situation. Another one is the permanent electrical power, bandwidth, and servers use required to support publication, as opposed to the one-off nature of printing publication. But even when the naming space and the publication resources are secured, organization and design of information can play a significant role in the resilience of resource on servers (Berners-Lee 1998). As Berners Lee, the inventor of the Web puts it:

> There are no reasons at all in theory for people to change URIs (or stop maintaining documents), but millions of reasons in practice. (Berners-Lee 1998)

Change of people, internal organization, projects, Web server technologies, naming practices, etc. can result in restructuring and sometime loss of information.

The growth of content management system (CMS) style of publishing gives, from this point of view, the illusory impression to bring order in chaos as CMS usually have one unified information structuring style and often archiving functionalities. The problem is that they add another layer of dependency on software (the CMS software), as no standardization exists in this domain. Information architectures based on CMS prove to be "cool" as long as the CMS is not changed, that is, not very long.

But whether information design is hand- or system-driven, the Web is not and would not become a self-preservation medium. The more fundamental reason is to be found in the contradiction between the activities of publishing and preserving. Publishing means creating newness even when it is at the expense of the old (in a same naming space for instance, as well as new and old books have to cohabit in the same publisher's warehouse). The experience proves that the incentive to preserve, is not sufficient among content creator themselves, to rely on them for preservation. Actually, the first step for preservation is to have it done by a different type of organization, driven by different goals, incentives and even a different

ethic. The Web as an information infrastructure cannot solve what is mainly an organizational problem. Therefore, archiving the Web is required as an activity made independent from publishing.

1.2.1.3 An Impossible Task?

Finally, the third category of arguments against Web archiving is supported by people acknowledging the need to archive the Web but skeptical about the possibility of doing it. Their skepticism is grounded either on the size of the Web itself, or on the various issues (privacy concerns, intellectual property, and copyrights obstacles) that challenge Web archiving.

The first aspect, the alleged immensity of the Web, has to be considered in relation to storage costs and capacity of automatic tools to gather huge amount of information. Current DSL lines and personal computer's processing capacity give the ability to crawl millions of pages every day. The scale of Web archiving means is in proportion with the scale of the Web itself. Even if the latter is difficult to estimate precisely (Dahn 2000; Egghe 2000; Dobra and Fienberg 2004), we know from different sources[2] that the size of the surface Web is currently in the range of tenth of billions pages, and that information accessible behind forms and other complex Web information system that cannot be crawled (the hidden Web) is one or two orders of magnitude larger (Bergman 2001; Chang et al. 2004). Archiving the surface Web has proven to be doable during an entire decade by the Internet Archive, a small organization with small private funding (Kahle 1997, 2002). The reason for this is that for the same amount of content, creators pay huge value for creation, maintenance, and heavy access. Storage is only a modest part of the cost of Web publishing today. The Internet Archive on the contrary, pays only for storage using compression (as crawl is donated by Alexa), and access, the latter being, per unit of content, much smaller than that of the original server. This results in the tangible possibility to host a quite extensive copy of the Web in a single (small) institution (see Chap. 9).

The second aspect, privacy concerns, intellectual property and copyrights obstacles would not be addressed in detail in this book.[3] Let us just note that the Web is primarily a noncommercial publishing application of the Internet. Private communications are not supposed to occur on the Web

[2] The sources are the documented size of search engines index (Yahoo claims to index 20 billion pages, Google says it index more (Battelle, 2005), the size of Internet Archive collection snapshots (10 billion pages)), recent studies based on sampling methodologies (Gulli and Signorini, 2005).

[3] Brown (2006) addresses these issues in more details.

but on communication applications (like the mail or instant messaging) and when they do (Lueg and Fisher 2003), there is always the possibility (widely used) to protect them by login and password. Spaces hence protected are not considered as part of the public Web and therefore should not be preserved in public archives. This natural delineation of the public/private sphere on the Internet is reinforced by the way crawlers operate (by following links) which means that pages and sites need to have a certain degree of in-linking to be discovered and captured. Others are disconnected components of the Web (Broder et al. 2000) that will naturally be excluded from crawls. One can also use this and set higher thresholds for inclusion in collection (more than one in-link) to limit capture to the more "visible" parts of the Web.

With regards to legal status of Web archiving, there are obviously various situations in each country and this is an evolving area. It is beyond the scope of this book to cover these aspects that have been addressed in Charlesworth (2003). Let us just note that the content published on the Web is noncommercial, either paid by advertisement on sites or paid by subscriptions. For all cases, Web archives, even with online access, have to find a nonrivalrous positioning with original websites and this can be done by respecting access limitations to content (as stated by the producer in robots.txt files for instance), having an embargo period, presenting less functionalities (site-search, complex interactions) and inferior performances (mainly speed access to content). Using Web archive to access content is thus done only when the original access is not possible and revenue stream, if any, for the original publisher is not threaten by Web archives (see on this topic Lyman 2002). On the contrary, Web archive can alleviate significantly, for site creators, the burden of maintaining outdated content and allow them to focus on the current. Even in this situation, authors and publishers may also request that their material be removed from publicly available archives. Request can also come from third-party for various reasons. How shall public Web archives respond to these requests?

Some recommendations have been proposed in the context of the United States, see Table 1.1 (Ubois 2002).

Table 1.1.

Type of request	Recommendation
Request by a webmaster of a private (non-governmental) website, typically for reasons of privacy, defamation, or embarrassment	1. Archivists should provide a "self- service" approach site owners can use to remove their materials based on the use of the robots.txt standard 2. Requesters may be asked to substantiate their claim of ownership by changing or adding a robots.txt file on their site

	3. This allows archivists to ensure that material will no longer be gathered or made available 4. These requests will not be made public; however, archivists should retain copies of all removal requests
Third party removal requests based on the Digital Millennium Copyright Act of 1998 (DMCA)	1. Archivists should attempt to verify the validity of the claim by checking whether the original pages have been taken down, and if appropriate, requesting the ruling(s) regarding the original site 2. If the claim appears valid, archivists should comply 3. Archivists will strive to make DMCA requests public via Chilling Effects, and notify searchers when requested pages have been removed 4. Archivists will notify the webmaster of the affected site, generally via e-mail
Third party removal requests based on non-DMCA intellectual property claims (including trademark, trade secret)	1. Archivists will attempt to verify the validity of the claim by checking whether the original pages have been taken down, and if appropriate, requesting the ruling(s) regarding the original site 2. If the original pages have been removed and the archivist has determined that removal from public servers is appropriate, the archivists will remove the pages from their public servers 3. Archivists will strive to make these requests public via Chilling Effects, and notify searchers when requested pages have been removed 4. Archivists will notify the webmaster of the affected site, generally via e-mail
Third party removal requests based on objection to controversial content (e.g. political, religious, and other beliefs)	As noted in the Library Bill of Rights, "Libraries should provide materials and information presenting all points of view on current and historical issues. Materials should not be proscribed or removed because of partisan or doctrinal disapproval" Therefore, archivists should not generally act on these requests
Third party removal requests based on objection to disclosure of personal data provided in confidence	Occasionally, data disclosed in confidence by one party to another may eventually be made public by a third party. For example, medical information provided in confidence is occasionally made public when insurance companies or medical practices shut down These requests are generally treated as requests by authors or publishers of original data

Requests by governments	Archivists will exercise best-efforts compliance with applicable court orders Beyond that, as noted in the Library Bill of Rights, "Libraries should challenge censorship in the fulfillment of their responsibility to provide information and enlightenment"
Other requests and grievances, including underlying rights issues	Other requests and grievances, including underlying rights issues, These are handled on a case-by-case basis by the archive and its advisors. Control, and reinsertions of Web sites based on change of ownership

This recommendation could be adapted in other legal environments while re-using the main practical mechanisms proposed (communications from the owner of the site through the use of the widely used robots.txt standard and alignment on what has been done on the original site for third party claims).

There is obviously a need for better understanding of the symbiosis between site creators and Web archives, that can, while respecting creator's rights ensure a memory can be preserved, but this is also a part of the maturation process of the Web as medium.

In sum, argument against the necessity as well as the possibility of archiving the Web appear unsurprisingly, in our view, to be inconsistent with the central role that the Web has in today's creation and diffusion of culture as well as with its sheer nature. Chapter 2 will provide further insight on how important Web archives are for research in many domains (Jones and Johnson 2006). We, throughout this book, try to demonstrate that, while posing serious challenge to traditional practices, Web archiving is both possible and one of the main items on the agenda of cultural heritage preservation today.

1.3 Web Characterization in Relation to Preservation

The Web has important characteristics that any preservation effort has to take into account. We review them in this section on different angles. The first one is the Web's cardinality, that is how many instances of each piece of content exists, the second is the Web considered as an active publishing system and the last one is the Web as a global cultural artifact, its hypermedia and open-publishing nature.

1.3.1 Web's Cardinality

The first question to address in cultural artifacts preservation is cardinality that is, the number of instances of each work that are being dealt with. Archives and Museum usually handle unique artifacts, even if in some cases, there are several casts, replicas or proofs for of a single sculpture, painting, or photo work.

Libraries, on the contrary, are mostly keeping nonunique items in their printed collection (manuscript preservation is closer to archive's practice from this point of view). Uniqueness has a deep symbolic and social importance (Benjamin 1963). It has also an obvious impact on preservation's practices. Libraries have always had a second chance to find printed books long after their publication. It has been estimated that more than 20 millions of books for 30,000 editions have been printed between 1455 and 1501 (Febvre and Martin 1976) which means that, on average, incunabula's cardinality was over 650. This cardinality entails that preservation can take place with a certain delay after publication as multiple copies will survive for a period of time, even in the absence of active preservation. It also occasions a natural level of redundancy in the system that libraries form together. Using the data from one of the largest bibliographic database (WorldCat) (Lavoie and Schonfeld 2005) find a three tier distribution of print work's cardinality in the libraries that use WorldCat (20,000, mostly in North America): 37% are held once only, 30% are held 2–5 times and 33% are held more than five times.

Time and redundancy are two significant advantages from a preservation perspective that reinforce each other. They have not always existed. Reproduction of manuscripts went on with imperfections for centuries before the invention of printing, therefore presenting, even when several (actually few) copies existed, variations. Librarians form the largest ancient library, the library of Alexandria, used to make copies of manuscripts that transited into the city, but they kept the original (Canfora 1989). Compilation, comment, annotation were often the main rationale for reproduction of text rather than authentic preservation, which added to the inevitable loss that manual copy entailed. More systematic copying of texts was often made for external reasons like when Greek texts have been systematically copied at the occasion of the invention of a new writing (the minuscule) in the Byzantine Empire in the ninth century, fixing and transmitting them in the form that we know today.

The coming of printing changed significantly the situation to this regard. It stabilized content while permitting its wider distribution (Febvre and Martin 1976; Eisenstein 1979). It also permitted, by augmenting significantly the cardinality of works, unprecedented preservation efficiency.

Where it has been estimated that 1 out of 40 known works from antiquity has been preserved (and less if we take in account unknown works),[4] preservation efficiency raised to more than 1 out of 2 in the seventeenth century in France and close 80% the century after (Estivals 1965), for one single institution, the Royal Library after the enforcement of a legal deposit by François 1er in 1537 (Estivals 1961; Balayé 1988). Today, preservation of printed works has achieved maturity and efficiency in most countries, both from the practical and institutional point of view, permitted by printed material's stability as well as cardinality.

Whatever it was, cultural artifact's cardinality was at least unified from creation to access. This is not any longer the case with the Web. Web's content cardinality is not simple but compound. As content's source is usually a unique server, one can sensibly argue that its cardinality is, like art works and manuscripts, one. It indeed presents the same vulnerability, even enhanced by the fact that content depends on the producer himself for its existence. But on the other end, access as well as copies of Web content can be virtually infinite. This gap between the two Web cardinalities leads us to the important notion of Web resource. A resource has a unique source (the Web server) and a unique identifier, but can be generated virtually infinitely and undergoes some degree of variation for each of its instantiations. From a preservation point of view, a resource has two important characteristics.

The first one is that it permanently depends on its unique source to exist. This makes a huge difference with printing where print masters are needed once only, after what, books live their own existence. The second one is that Web servers can tailor content for each instantiation of the resource, making it different each time for the same URI. The Web is, from this point of view, not a container with fixed files, but a black box with resources, of which user only get instantiations.

As Krishnamurthy and Rexford explain about the Web protocol:

> One way to understand the protocol is to imagine that the origin server contains black boxes representing resources denoted by URIs. An origin server applies the request method to the resource identified by the URI and generates a response. The common understanding of reading a resource from a file and writing the response back to the client is abstracted away in the black box view. This view generalizes the notion of a resource and separates it from the response sent to the client. Different requests for the same URI can result in different responses, depending on several factors: the request header fields, the

[4] Hermann Strasburger cited by Canfora (1996).

> time of the request, or changes to the resources that may have happened. (Krishnamurthy and Rexford 2001)

The fact that Web preservation is dealing with resources, with the paradoxical cardinality that this entails, has several implications. The first one is that, given that a virtually infinite number of copies can be made easily, one can have the misleading impression that active Web archiving is not required for preservation. However, the multiplicity of instantiations hides the extreme dependence on one single source (the server) that can be removed, updated, etc. at any time, thus the need for an active archiving.

The second implication is that Web archives can only capture some instantiation(s) of resources, with, potentially a degree of variation amongst them.[5] This is the case when the content is tailored for a specific browser, a certain time or a certain geographic location or when the content is adapted for each user.

As we will see in the next section, the Web is indeed an active publishing system, and therefore variance of responses is actually an important aspect to consider when archiving.

1.3.2 The Web as an Active Publishing System

The Web is the main publishing application of the Internet. As such, it consists mainly of the combination of three standards, the URI (Berners-Lee 1994) defining a naming space for object on the Internet,[6] HTTP (Fielding et al. 1999) defining a client–server interaction protocol using hyperlinks at its core, and HTML (Berners-Lee and Connolly 1995) an SGML DTD that defines the layout rendering of pages in browsers. The implementation of these three standards enables any computer connected to the Internet to become a publishing system. Together, the network of Web servers forms a

[5] Dynamic generation of pages is also used for unifying design and architecture of information (navigations devices, etc.) across the entire site. The use of templates makes it easy for pages to look alike and eases the change of design by change of template(s) rather than individual pages. It has been estimated that templates based pages represent 40–50% of all pages (Gibson et al. 2005). Eiron and McCurley (2003) also found on a billion page crawl, 40% of pages including a "?" character in their URL, which is usually used to send a query to a database and generate dynamic pages.

[6] This standard is considered as being the most important of the three by the inventors of the Web (Berners-Lee and Fischetti 2000; Gillies and Cailliau 2000) as it positions the web as a universal access overlay on any documentary resource accessible on the Internet.

unique information system that can be used to generate, update and publish content in any manner that modern computing allows.

Compared to other publishing medium, it represents a revolution in publishing, extending possibilities in all possible directions for generation, organization, access, and rendering of content. Consider for instance linking: one can argue that this is just a modern form of reference that existed already since the earliest times of writing.[7] But the fact that it is actionable on the Web changes the way references are used by fragmenting content to smaller addressable pieces and overall favoring transversal navigation and access to content which in return, deeply changes the nature of writing as well as reading (Aarseth 1997; Landow 1997; Bolter 2001).

The fact that content only exists on the system, and more precisely on the publisher's servers, makes content's existence dependent on permanent publishing from the creator. Whereas a book can live its life independently of its publisher after leaving the print workshop, Web content is granted no existence beyond its original server (at the exception of course of transient caching mechanisms (Hofmann and Beaumont 2005)). Permanent publishing extends dramatical control that creators have over content. With the Web, they can at any time change, update, and remove in real time items from "publication". Furthermore, Web producer are using Web information systems (WIS)[8] that can combine, aggregate and re-organize information from almost any type of exiting information system (database, document repositories, applications, etc.). The Web is therefore not a fixed information space but an active publishing space, resulting effect of a mingled set of active information systems.

Web archives hence need first to separate content from its original creator's permanent publishing and second, to ensure that content can be resilient from the current Web's failure and evolution. The former requires copying and archiving content in a separate infrastructure (see below and Chap. 3, Roche 2006). The latter entails active preservation of web content (see Chap. 8 Day 2006) to remove dependency from the various system's components (protocols, digital formats, applications, etc.) and their inevitable technological obsolescence. Web preservation shares this need for active technological preservation with digital object in general, but the

[7] For a comparison to traditional scientific citation and how it can be used to 'measure' science see Ingwersen (1998), Björneborn and Ingwersen (2001). See a critical analysis of this possibility in Thelwall (2001), Thelwall and Harries (2004) and Thelwall (2006).

[8] On Web information systems see from a system perspective (Florescu et al. 1998; Antoniol et al. 1999; Scharl, 2000) and from a user and designer's perspective (Rosenfeld and Morville, 2002).

separation from the creator's permanent publishing is specific to Web preservation.

But removing any dependency from the original server entails that, from the various functionalities and mode of interaction that the Web offers, Web archives can only preserve a few. There is a cost for the separation from the network of original Web information systems.

Functionalities that are executed on client-side are the ones that one can reasonably ambition to preserve. The range of functionalities embedded in the page's and related file's code executed by the client will, most of the time, be executable on the archived versions, but functionalities provided by server-side code and/or information will not. It is still possible to document aspects of the original material that are lost (like specific types of interactivity that one can record on video), but this can only be done for a limited number of pages, a certain point of view and a specific situation (Christensen-Dalsgaard 2001; Brügger 2005).[9]

1.3.3 The Web as a Cultural Artifact

In addition to being an active publishing system, the Web is also information space with its own specificities. The word "Web" designate in this context a vast digital cultural artifact (Lyman and Kahle 1998) that can be characterized by the fact that:

- It is published and available (mostly freely) from any place connected to the Internet
- It is structured as an hypermedia using direct and actionable links between content pieces[10]

[9] Interesting also is the website designer's viewpoint on this issue. In Dubberly et al. (2002) Challis Hodge, suggests to archive for sites he is designing:
- Request for Proposal (RFP);
- Statement of purpose and intended use;
- Description of context of use (examples as needed);
- Description of the actual and intended users;
- Static representations which adequately capture overall look and feel;
- Examples of several key paths through the site;
- Description of underlying and supporting technologies;
- Any relevant modules such as flash animations, movies, PDFs, etc.

[10] Eiron and McCurley (2003) show that a third of links extracted from a billion-pages crawl point to the same directory, a third link across, up or down in the hierarchy of directories from the same site, and a third link to external sites.

– It contains not only text but any combination of images, sounds and textual content
– It is the result of a distributed and open authorship[11]

Although the Web does, to a large extend, re-use previous forms of publishing[12] (Crowston and Williams 1997; Eriksen and Ihlström 2000; Shepherd and Polanyi 2000), it also invents new ones. This is the case for instance with blogs that combine an extreme simplicity to publish (even technical skills that were required for normal sites are not any longer required), a powerful reference management (including reverse reference or citation notification using ping back) and facility to update, add comment and remove content, all this resulting in the open publication of personal comments by tenth of millions of people.[13]

This characterization of the Web as a distributed hypermedia openly and permanently authored at a global scale entails that Web archiving can only achieve preservation of limited aspects of a larger and living cultural artifact.

The interconnectedness of content is a major quality of the Web that raises issues when it comes to archiving. This issue is discussed further in Chap. 3 (Masanès 2006b), but as a general consequence, it appears that archiving always implies some sort of selectivity, even if it is not always in the sense of manual, site-by-site, selection. This argues for large and broad archiving to avoid, as much as possible, to cut the information continuum that the Web represents (Lyman et al. 1998) or to the definition of a specific analytical purpose for grounding selection decisions (Brügger 2005). But practically, crawl implementation in terms of priority and policy (see section on "Client-Side Archiving") or manual selection, involve that the archived portion of the Web will always only be a slice in space and time

[11] Authorship is no longer limited to a few people but is distributed across tenth if not hundreds millions people. It can been estimated in the case of France for instance that publication nodes, that is, person or structure that publish (editors not authors), have been extended of three orders of magnitude from printing to the Web: from around 5,000 publishers or structures of publication to more than five millions site and personal sites (source: Association Française des Fournisseurs d'Accès Internet). This does not include weblogs.

[12] But this was the case for printing as well, which for a long period did imitate manuscript writing and page organization before inventing its own (Febvre and Martin 1976).

[13] On blogs preservation see Entlich (2004).

of the original Web. How to make this sampling meaningful and representative of the larger Web? What implication will this have on future understanding of what the Web was? All these questions have to be considered when engaging with Web archiving. Even the definition of what the "original Web" is, raises issues as it is the collective experience of Web users or the totality of instantiations of content that we should, as seen before, consider rather than a pre-existing set of fixed content.

Another characteristics that obliges to re-conceptualize and re-organize traditional preservation practices is the open authoring nature of the Web. It indeed makes it very difficult to filter and structure preservation based on publishers and authors. They are just too numerous on the Web and they are difficult to identified and register. Sometime, authorship information is available on the site, sometime not, and sometime not in a reliable way. The only information registered (in a quite loose and uncontrolled way) is information about who rents the domain name for the DNS management. Although this information is certainly of great value to complement archived Web material, it is certainly not easy to interpret and use directly.

As a cultural artifact, the Web thus presents a different style of information organization and therefore different structural patterns to use for its preservation. Trails of content linking and users' navigation form the natural structures that archives have to use most of time to organize their gathering for instance. The Web's characteristics require hence deep transformations of preservation methods. Holistic approach to Web archiving is more prone to adaptation to the Web's characteristics, but any type of Web archiving should incorporate them at the core of its method.

1.4 New Methods for a New Medium

Libraries, archives, museums have long developed very efficient methods adapted to their holdings that have played a key role in the building of society's memory. Although much has to be learned and can be reused for Web preservation, the Web's nature and qualities require, as we have seen, to re-think and adapt preservation practices inherited from this long tradition of physical cultural artifact preservation. This section will present an overview of the new methods and approaches that have to be used for Web preservation (Chaps. 3–8 provide detailed discussion of most of them).

1.4.1 Web Preservation and Information Infrastructure

Before starting the methodological discussion, we need to address first the larger question of Web archive's positioning in the information infrastructure[14] in general and the Internet in particular

Borgman (2000) discusses the definition of a global digital library (pp 47 sq.) and explains the difference between an evolutionary and revolutionary view of the information technology:

> "The revolutionary view is that digital libraries are databases linked by computing networks, and that, taken as a whole, they can provide an array of services that will supplant libraries. The evolutionary view is that digital libraries are institutions that will continue to provide content and services in many forms, just as their predecessors' institutions have done, and that eventually they will supplement libraries, as they exist today" (ibid. p. 48)

She herself proposes a middle-ground definition of co-evolution that states: "digital libraries are an extension, enhancement, and integration both of information retrieval systems and of multiple information institutions, libraries being only one (…). The scope of digital libraries' capabilities includes not only information retrieval but also creating and using this information".

The situation for Web archives is different in the sense that they are dealing with an existing and already structured information space, which is also openly accessible. By design, there is no need, in this space, for gatekeepers as there are no physical access limitations. To this regard, the role of Web archives is, by nature, more modest in term of information organization.

---

[14] The concept of infrastructure in general is defined in Star and Ruhleder (1994) with several dimensions:

- Embeddedness;
- Transparency;
- Reach or scope (infrastructure has reach beyond a single event or one-site practice;
- Learned as part of membership (new participants acquire a naturalized familiarity with its objects as they become members);
- Links with conventions of practice;
- Embodiment of standards;
- Built on an installed base;
- Becomes visible upon breakdown;

This concept is discussed in the context of information infrastructures in Borgman (2000, 2003).

Physical libraries had to create both a physical and intellectual organization of objects, and this allowed a large range of possibilities and choices. They had also, as they managed physical access to content, an unavoidable intermediary role. Digital libraries are extending this intermediary role by creating collaborative and contextual knowledge environments beyond the basic function of search and access (Lagoze et al. 2005).

Web archives, on their part, are dealing with content loaded with embedded and actionable relations and rich informational structures created by millions of people globally editing the Web. When traditional archives and libraries are providing their own organizational view and tools on this content (in subject gateways, Webographies, etc.) they are only participating in this global editing of the Web. This does not diminish their inner qualification as domain experts but it positions it in a larger organizational effort.

As Web archives, they have more responsibilities, as they will capture and freeze both content and context and they can have the temptation of recovering their ancient unique role of information organizers. But in doing so, they can only achieve to freeze and preserve their own sample of a larger living cultural artifact. This can be legitimate when grounded on selection policy fitting a community of users' need or driven by clearly define research goals (see Chaps. 3 and 10 on these issues (Lecher 2006; Masanès 2006b)). But the costs and limitations of doing so, as well as the technical possibility to archive both at a larger scale and in a more neutral way require considering also an alternative way of archiving the Web. This alternative is more modest in role but more ambitious in scope. The role of information organizers is limited to capture and be faithful to the original structure generated by millions of people globally editing the Web. Exhaustiveness being out of reaches, as we have seen in the earlier section, one can at least have the ambition of neutrality in capture of content, by following the distributed and collective nature of the Web to guide the capture and extend it as much as possible along these lines. Ambition is thus on quantity, and this is merely a matter of scaling up technical resources. This has been the approach of several national libraries for their national domain and the Internet Archive at a global scale.

None of these initiatives can provide alone extension, depth and quality of content archived. The various efforts will be considered as part of one global archive when interconnection between the Web archives will be organized as interconnection between publishing servers is through the Web. Only this will enable users to leverage all these efforts and result in the best Web memory possible. In this sense, the larger the participation of different institutions and individuals, the better as they can complement each other and offer different angle, depth and quality of archiving. But

this requires that they become part, at some point, of a larger Web archives grid. Such a grid should link Web archives so that they together form one global navigation space like the live Web itself. This is only possible if they are structured in a way close enough to the original Web and if they are openly accessible. The International Internet Preservation Consortium (IIPC) has been working on laying the grounds for the former by developing standards and tools that facilitate building this type of archiving (some will be described in the rest of this section). Open access is a matter of regulation and policy and remains at this point in time an open issue.

Web archives, individually or as a whole, can fit naturally in the existing Internet infrastructure. They are using the same protocols and standards for organizing information and providing access to it. The Web can naturally include them as they are entirely compatible with it.[15] From the infrastructural point of view, Web archives can hence easily find a position as a complement of the existing Internet infrastructure. They are providing a Web memory that is part of the Web itself and limits the negative impact of the necessary transient nature of Web publishing.

One could be unsatisfied by the modesty of this role, but this would be neglecting the value of the distributed and collective nature of this medium that justifies it.

1.4.2 Acquisition

The term "acquisition" designates the various technical means used to get the content into the archive. This includes online capture as well as off line delivery of content. It does not cover the selection neither the ingest process with metadata generation.

From the technical point of view, this interaction phase with the producer, traditional for memory institution, is anything but trivial in Web archiving. The reason is that no single-approach suffices to cover efficiently the wide variety of Web publishing techniques. The widening of producers range and the increasing size of content is to a certain extent balanced by the automation made possible in the Web environment. However, the main obstacle that acquisition tools have to overcome is the HTTP protocol inability to provide bulk copy of server's content. HTTP servers can only deliver their content file by file, as long as their URI are requested. This makes the discovery of individual path to each file one of the key issues in Web archiving.

[15] One could argue that there is here a potential infinite regression, to what it can be opposed that web archives should avoid archiving other web archives, and limit themselves to the live web.

We review in this section the three types of acquisition methods. Why three methods? Mainly because the gathering process can either be done remotely as client, close to the output of the server or by direct access to the server's files. The first option is made with archiving crawler or website copier, derived and adapted from search engine technology, provide a powerful tool for capture in the position of client. Chapter 4 (Roche 2006) gives a detailed description of theses tools and their application for Web archiving. We will only present in this chapter an overview of this technology that permits to evaluate in which case it can be applied. As the crawler is, for the Web server, a client like any other, we use the term "client-side archiving" for this acquisition method. Depending on the Web server backend architecture and level of interaction with the client, crawlers can capture either the full website, or some portions of it only. The portion left out of reach for crawlers have been called "deep Web" or "hidden Web" in the search engine terminology. We will endorse this terminology as long as it remains clear that the delimitation of the hidden Web is purely technical and continuously moving as crawlers improve their ability to find path to documents.

Two alternative methods exist to gather content even if they have been far less applied and remain even investigational so far. Both need to be operated from the server side, which requires not only an authorization but also an active participation of the site publisher to be used. The first one is based on users of the site, exploiting their navigation path and exploration of the site's content to archive it. As it is based on the recording of transactions made between users of a site and the server, we call it "transaction archiving". The second consist in archiving directly from the publisher the various component of his or her Web information system and transform them to an archival form. It is called accordingly, "server-side archiving". These alternatives techniques are more demanding than the client-side archiving because they require, as mentioned above, an active participation from the producers but also because they have to be implemented on a case by case bases. But even if they do not scale up, they can be applied in cases where crawler fails to capture accurately and when the content deserves it. A detailed technical presentation of the crawlers limits and alternatives techniques for archiving the hidden Web can be found in Chap. 5.

1.4.2.1 Client-Side Archiving

This is the main acquisition method both because of its simplicity, scalability and adaptation to a client–server environment (see Fig. 1.1). Crawlers are adapted to what is the usual way of accessing to the Web. This allows archiving of any site that is accessible either freely on the

open Web, either on intranets or extranets, as long as the crawler get the appropriate authorization. This method not only adopts the same position as normal Web users, it also imitates its form of interaction with servers. Crawlers start from seed pages, parse them, extract links and fetch the linked document. They then reiterate this process with document fetched and proceed as long as they have links to explore[16] and they find document within the scope defined. This process is needed, as HTTP does not provide a command that would return the complete list of document available on the server, contrary to FTP for instance. Each page has therefore to be "discovered" by link extraction from other pages.

The crawling technology has originally been developed for indexing purposes.[17] Application to Web archiving, despite the fact that is re-use most aspect of this technology implies several changes to it.

The first one is that archiving crawlers shall try to fetch all files, whatever their format to archive a complete version of sites, contrary to search engine crawlers who usually fetch only files they can index. Search

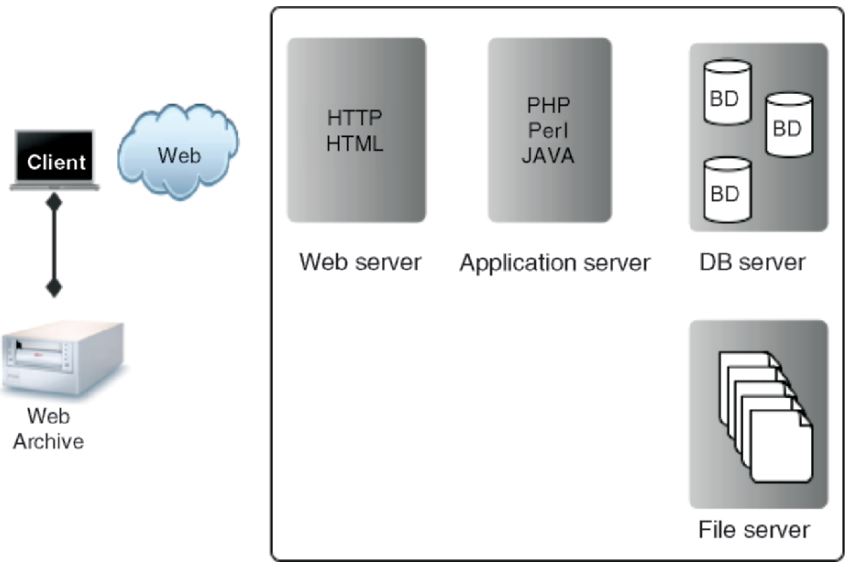


Fig. 1.1. Client-side archiving: the Web archive is in position of client to gather content from the Web server. The Web server can generate content from various other servers (application, database, file servers)

[16] For recent overviews of crawling technology see: Pant et al. (2004) and Chakrabarti (2002).

[17] For an overview of commercial search engine development see: Sonnenreich (1997).

engines crawlers for instance often ignore large video and application files Downloading this type of files can make a significant difference in term of time and bandwidth needed for crawling entire sites.

The second difference is related to temporal management of crawls. For avoiding overload of Web servers, crawlers respect politeness rules (imposing a fix delay between two requests, usually several seconds, or a delay that depends on response time from the server, see Chap. 4 for more details on this topic (Roche 2006)). This entails that a Web capture can span during several minutes at best, several hours and sometimes several days. A simple calculation shows that when respecting a delay of 3 s between two requests, it will take more than three days to archive a site with 100,000 pages. This delay raises the issue of temporal consistency of the capture as site can undergo changes during the time they are being captured. If the index page is changed during the capture for instance, its archived version will not be consistent with the more recent one that linked to the last pages archived.

This is an issue for archiving crawls because the crawl is supposed to provide content and not only direction to content. Search engine crawls are only used to point to live pages on the Web which means that hypertext context for them is the one provide by the original server (which is, of course, supposed to be consistent across pages and updates). On the contrary, archiving crawlers have to capture content as a whole, which will, with or without its internal coherence, remain as the only context for navigation and interpretation.

This has far reaching consequences with regard to crawling policy. As politeness to servers has always been a bottleneck for crawling, SE crawlers have been using mainly breadth-first crawling priority, with some variants mainly aiming at crawling "best" pages first (Cho et al. 1998; Najork and Heydon 2001; Najork and Wiener 2001; Castillo et al. 2004; Baeza-Yates and Castillo 2005). Adopting this policy is also a way of minimizing impact of robots traps on the overall crawl by laying out the crawl over a large number of different sites.

But this crawl scheduling strategy has the inconvenience of augmenting temporal discrepancy of crawls at the site level.

It has therefore been proposed to adopt for archiving crawls a site-first priority.[18] But, for large-scale crawls it is still necessary to optimize crawl efficiency by making sure resources are used at their maximum capacity. Given delay between requests and crawling resources available, one has to

[18] This was for instance discussed for the requirements of Heritrix within the IIPC (Masanès, 2004). On crawl scheduling policies that incorporate site as a vertical dimension see (Castillo et al. 2004; Baeza-Yates and Castillo 2005).

find the optimal number of sites to start at the same time to make sure request frequencies will be set by politeness rules, with no unnecessary delay between requests. Figure 1.2 shows the "front line" of a crawl, which size corresponds to the optimal allocation of crawl resources.

There are limits to what can be achieved using this method. Most occur during link extraction and some during retrieval through the HTTP interface. The former can be caused by the fact that URI extracted are badly formed or use complex parameters, by the difficulty to parse URI from scripts or executable or even HTLM code. The latter can be caused by re directions, negotiation of content, authorization, slow responses, extreme size, TCP connections anomalies, invalid server responses, etc. For more details, see Chap. 4 (Roche 2006), see also for a taxonomy of various issues in Boyko (2004). For a presentation of Heritrix, a large-scale archiving crawler that implements the frontline developed by the Internet Archive and the Nordic Libraries based on requirements of the IIPC, see Mohr et al. (2004).

Use of this type of tools allows large scale acquisition of content in a holistic way, that is not

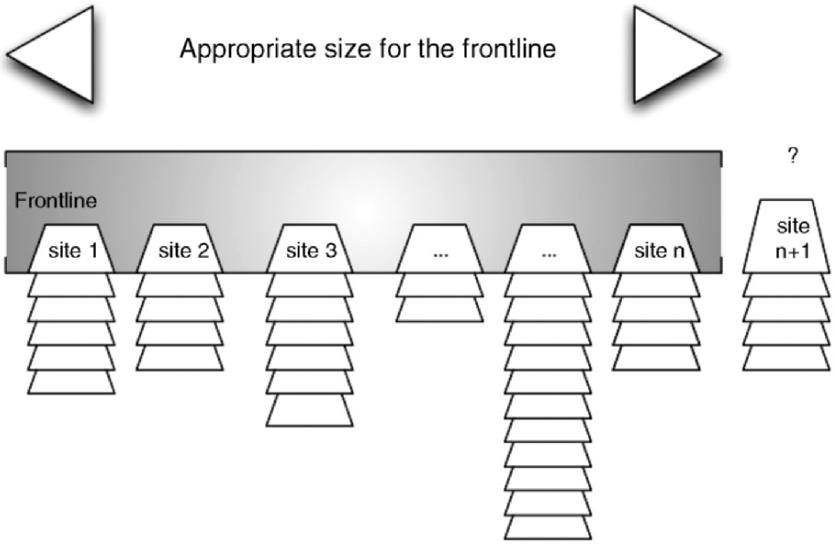
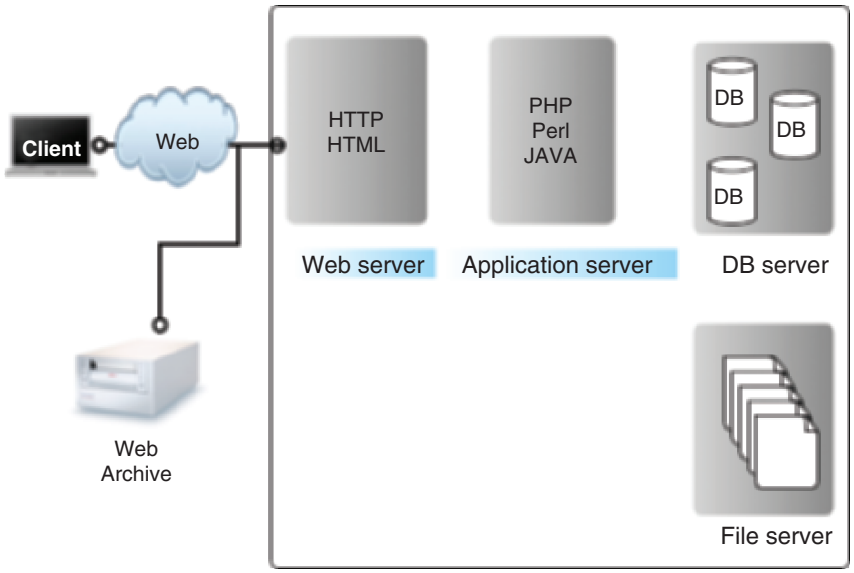


Fig. 1.2. The frontline contains sites to be crawled at the same time by the same crawling infrastructure. Size of the frontline (n) is optimized if delay between requests is limited by politeness rules only and crawling resources are kept busy. If $n + 1$ sites were crawled, crawling resources limitation would introduce an additional delay and temporal incoherence. If $n-1$ sites were crawled at the same time, resources would be underused

1.4.2.2 Transactions Archiving



Transaction archiving (see Fig. 1.3), proposed by Fitch (2003), consists in capturing and archiving "all materially distinct responses produced by a website, regardless of their content type and how they are produced." This is implemented in the PageVault[19] system by using a filter into the Web-server's input (request) and output (response) flow. This functionality is now also available on some web content management systems like Vignette TM.

Unique request/response pairs are stored and archived, thus creating a complete archive of all content viewed for a specific site. Requests with only slight ("nonmaterial") differences are considered as unique by excluding from the calculation of the checksum the portion of code that codes them. How exactly this can be adapted to the numerous way of personalizing content is not clear.

This type of Web archiving can certainly prove useful to track and record every possible instantiation of content. Content never viewed will not be archived (as mentioned earlier, Boufkhad and Viennot 2003, have estimated that 25% of pages of a large academic website where never accessed). But hidden Web content, as long as it is accessed, will be recorded, which is a significant advantage.

[19] http://www.projectComputing.com/products/pageVault

The main constraint of this method is the fact that it has to be implemented with agreement and collaboration of the server's owner. It is therefore indicated mainly for internal Web archiving. It has the advantage to enable recording of exactly what was seen and when. For corporate and institutional archiving, often motivated by legal accountability, this can be an advantage. It is even possible to combine this with information from the log server, about who did view the content. Obviously, what can be seen as an advantage for internal Web archiving, would be a problem for a public archive, as it could raise serious privacy concern. But it is not usable in this context anyway.

1.4.2.3 Server-Side Archiving

The last type of acquisition method for Web archives is to directly copy files from the server, without using the HTTP interface at all. This method, as the previous one, can only be used with the collaboration of the site owners (see Fig. 1.4). Although, it seems to be the most simple, it actually raises serious difficulty to make the content copied usable. Even in the case of static HTML files, one would have some difficulty to navigate in

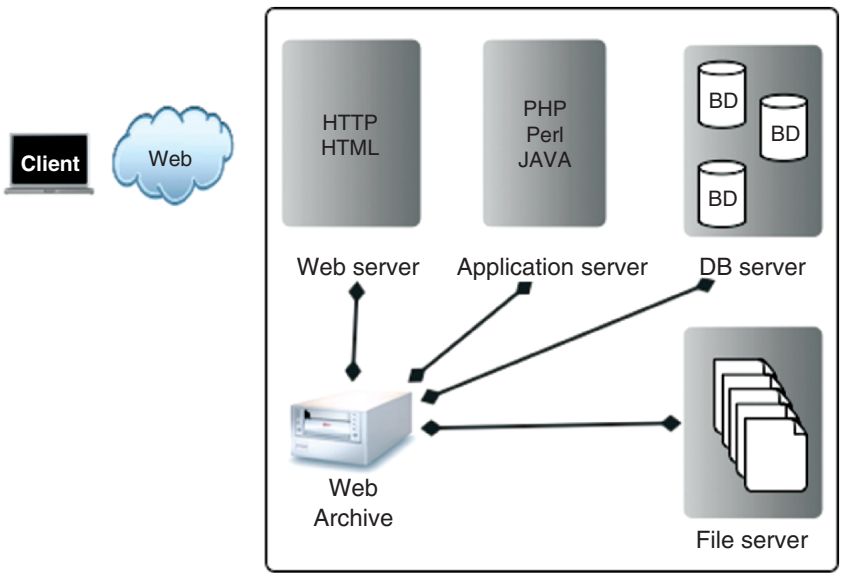


Fig. 1.4. Server-side archiving: different pieces of information are archived directly from servers. Generating a working version of the archived content and not only a back-up of files is the challenge of this method

the content through absolute links as the domain name will be different in the archive. But most of the problem comes from dynamically generated content, that is content aggregating pieces from various sources (templates, database) generated on-the-fly by user requests. Copying database files, templates, and scripts does not mean that it will be easy to regenerate content from the archive. On the contrary, it will certainly be a challenging task as it required being able to run the same environment, with the same parameters in the archive. Actually, when possible, dynamically generated content is better preserved in its final form, usually flat HTML files (this is the case for most CMS, blogs and wikis for instance).

But it is sometime difficult, even impossible for crawlers to find path to some documents of a website and files that can only be accessed through a complex interaction (like sending a query to a form) will hardly be captured by crawlers (see Chap. 5, section on "archiving documentary gateways", Masanès 2006a). This portion of the Web, called "hidden" or "deep Web" is estimated (Bergman 2001; Chang et al. 2004) to be larger than the "surface" Web (also called publicly indexable Web[20]).

In this case, server-side archiving can be a solution. As mentioned above, it requires active participation of the site administrator. More than a simple back-up which does not guaranty access to content in its original presentation, it implies being able to "play" again the site in the archive environment. This implies reducing dependency on database and server-side scripts execution as much as possible. This can be done by extracting the structured information contained in database and migrate it into XML. A typical information architecture called documentary gateway that contains non-Web documents with that are accessed by a catalog can be archived like this. This has been done for several sites that pertain to the category of hidden sites by the Bibliothèque nationale de France (see Chap. 5).

This was only possible in the framework of the legal deposit that applies in France like in many other countries. The fact is that the hidden Web is also often very rich contentwise as it is with this type of information architecture that pre-existing large mass of content has been published on the Web. The popularity of this type or information architecture, makes the server site archiving, a method to consider where it can be applied.

[20] This term is used to designate the portion of the web that can be indexed by crawlers (Lawrence and Giles 1998, 1999).

1.4.3 Organization and Storage

As we have already seen, making a copy of a Web site is a nontrivial task. It actually implies to recreate an information system that will be accessible for users. As Antoniol et al. (1999) put it "Web site may be as simple as a single file or one of the most complex collection of cooperating software artifacts ever conceived."

Ideally the archive could be isomorphic to the original (same hierarchical structure, naming of files, linking mechanism, format) but for practical reason, it is almost never the case. As seen in the precious section, the acquisition of sites induces in certain cases a transformation of files to be effective.

More challenging is the re-creation of the Web information system alike. WIS represent complex information architectures dependent on specific operating systems, servers configurations and application environment that would, in most cases, even be difficult to re-create from scratch for their designers and managers. This is the reason why Web archivists have to adopt transformation strategies. These transformations can impact addressing and linking mechanisms, formats, as well as object's rendering itself.

Three strategies have been adopted so far for structuring Web archives. The first strategy is to create a local copy of the site's files and navigate through this copy in a similar way as on the Web. The second one is to run a Web server and serve content in this environment to user's browsers. The third option is to re-organize documents according to different (non-Web) logic of naming, addressing and rendering. The following sections present the pros and cons of these different strategies as well as their preferred use-case.

1.4.3.1 Local Files System Served Archives

Description

This type of archive (see Fig. 1.5) is based on the possibility that the URI specifications offers to use the local file system prefix "file" in a URI scheme to copy and access locally files from the original website like in this example:

HTTP://www.example.org/example.HTML

file:///Users/archive2005/example.org/example.HTML

This enables the use of the local file system for navigation through archived Web material. It also requires using a partial (relative) form of the URI eluding not only the prefix but also the server's name and the path of the object.

Standard browsers can open directly (i.e., without a Web server) such locally stored files and, as long as links in documents are relative ones, navigation on the archive will be the same as on the original site, noticeable only in the address bar of the browser when looking at the URI prefix (here "file" instead of "HTTP").

Comment

The main benefit of this strategy is to simplify access to the archive by mapping the original website structure onto the archive file system. Using standard browser and file system allows avoiding extra overhead associated with running Web server-based access archive. Therefore, even team with very basic IT technical skills can set up and run this type of archive. But there are several limitations in this approach. From a conservation point of view, the main shortcoming is that several transformations of the original files are needed. Therefore, strict faithfulness to the original cannot be respected except by documenting carefully changes applied to the original, or/and by keeping a copy of the original. Transformation of content is required at two levels in "local FS" archive's approach.

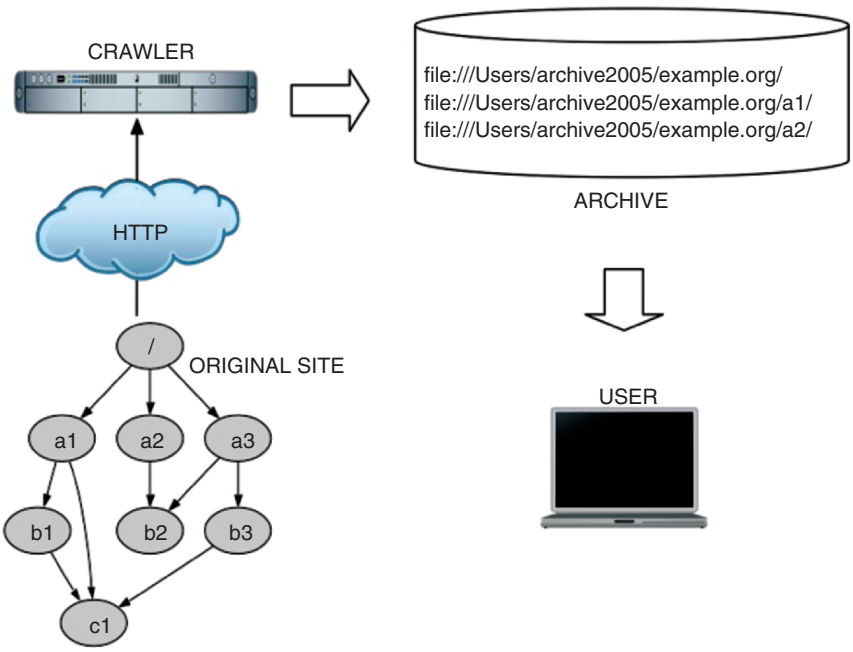


Fig. 1.5. Local file system archive. The original site is crawled and pages and other files are stored individually on the file system of the archive. Access is made by navigating directly on the file system

First, because of difference in the naming convention between URI and file system (allowed and reserved characters, escaping rules, case sensitiveness), the naming of objects may need to be changed (see Chap. I-B for a detailed presentation of these changes). In the case where the page is queried with parameters and generated dynamically, a name has even to be created for the resulting page, including the parameters to ensure uniqueness of the archived page.

Second, absolute links must be transformed in relative links in the page code itself to allow the file system-based navigation. Even if this can be documented simply by transforming the original URI into a comment in the code, this implies a manipulation of the original (see Chap. 4 for more details on these transformations).

From a practical point of view, the main shortcoming comes from the file system itself, a notably different architecture of information than the Web. First, the archive organization has to fit in the hierarchical organization of file systems. Yet, an archive is not only composed of sites but also of groups of sites (collections) and versions of sites. Mapping this organization to a hierarchical structure does not go without change and choices. How should sites be grouped together in a manner that resist time is a key issue to consider for instance. Collection names have to be persistent, time grouping have to be adapted to the capture frequency. On all these issues, thorough decisions have to be made beforehand. They will impact on how the chosen structure will persist as the collection develops. Organizing time transversal navigation (from one version of site to other) is also a key issue for which a layer of software has to be added on top of standard file system. This layer has to be able, at least to bind together different version of sites depending on their date (versioning) and present this to an appropriate user interface to navigate through time simply. This has often been implemented using an external management database of sites and captures information, and tools to generate intermediary presentation pages with a list of date at which the document has been archived.

An other limitation of this approach is due to the huge number of files Web archives have to handle. It is common to see archives with billions of files. This figure reaches the limits of current files systems capacity. Even when a FS can handle this amount of files, performance can be affected. To alleviate the load put on FS, large-scale archives have used container files. But this, of course, breaks the direct correspondence in naming and linking that the local FS archive's approach offers and entails to adopt the second approach, the Web-served archive (see below) to deliver content from these container files.

Preferred Use

This method is recommended for institutional or corporate site archiving and small scale nonincremental archiving. Depending on the use of this archive, the authenticity issue should be considered carefully, especially for institutional archiving. For small scale incremental archiving, the balance between difficulty for organizing persistently collection of files and the simplicity of access provided by this approach has to be appraised on a case-by-case basis.

For middle and large-scale Web archives, this method should be avoided.

Tools

This strategy is the simplest to implement for small and middle scale Web archive with many tools available like HTTrack for instance (see Chap. 4).

1.4.3.2 Web-Served Archives

Though more demanding, this option enables a better compliance to the original naming and structure of documents (see Fig. 1.6). It also permits to avoid file system size limitations, which is crucial for large scale Web archives.

Description

This method is based on response archiving (compared to the first one which is based on file archiving). Responses from the original server are stored unchanged in WARC container files[21] which permits to serve them back later to users of the archive with an HTTP server.

A WARC file records a sequence of harvested Web files, each page preceded by a header that briefly described the harvested content and its length. Besides the primary content recorded, the WARC contains also related secondary content, such as assigned metadata and transformations of original file. The size of a WARC file can vary up to hundreds of megabytes. Each record has an offset, which enables direct access to individual records (Web files) without loading and parsing of the all WARC files. Offsets of individual records are stored in an index ordered by URI. It is hence possible to rapidly extract individual records based on their URI out of a collection of WARC files, which is adapted to navigational access. The records are then passed to a Web server that provides them to the client.

[21] An earlier version of this format has long been used by the Internet Archive and is now standardized in a new version by the International Internet Preservation Consortium (IIPC) and has been submitted to ISO.

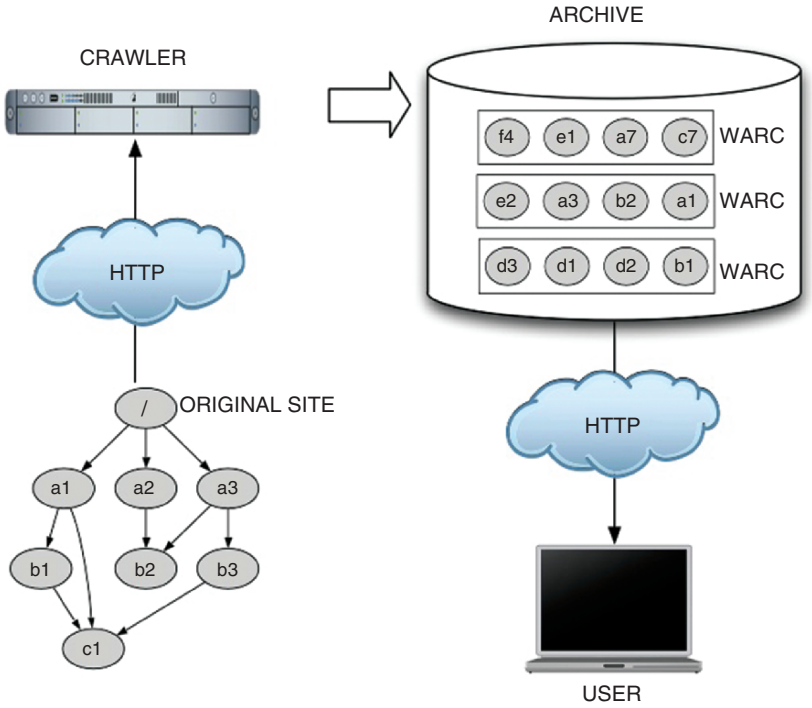


Fig. 1.6. The Web-served model: the original site is crawled and responses are stored unchanged in container (WARC files) which permits to avoid mapping to the file system's naming conventions and changing link structure. Access requires a Web server that fetches content in the containers and sends it as a response to the user

The conservation of the original naming scheme (including parameters in dynamic pages) allows navigation in the site as it has been crawled. The archive user can traverse all the paths followed by the crawler again.

Comment

The main advantage of using WARC containers is the possibility of overcoming the storage file system limitation in term of size (fewer individual files are eventually stored in the archive's file system) and namespace (the naming of individual Web files can be preserved). The Internet Archive achievement through the Wayback Machine (that gives access to 500 tb of Web collections) shows that this approach scales up like no other (see Chap. 9, Kimpton et al. 2006).

The downside of this approach is that direct access to the stored files is impossible. Two extra-layers of application are necessary to access content: a WARC file index system and a Web server (on this type of access, see

Chap. 6, (Hallgrímsson, 2006). These two layers are not outstandingly complex but require a running access environment, which can be difficult to set up and maintain in small organizations. This mediation can also raise problems for content rendering, as it requires that linking mechanism be appropriately mapped from the live–Web environment to the archive environment (we assume that original links have be kept unchanged in the archive, which is the main benefit of this method). This can be done at the page presentation level and at the proxy-level.

The first option consists in adding to the page sent to the archive user's browser a script that will, on the fly re-interpret links in the page to point to the archive (or change them in a relative form). The Internet Archive for instance does this with the following Java-Script code appended to each page sent to users.

```
<SCRIPT language="Javascript">
<!–

// FILE ARCHIVED ON 20050308085053 AND RETRIEVED FROM THE
// INTERNET ARCHIVE ON 20060514055212.
// JAVASCRIPT APPENDED BY WAYBACK MACHINE, COPYRIGHT
INTERNET ARCHIVE.
// ALL OTHER CONTENT MAY ALSO BE PROTECTED BY COPYRIGHT
(17 U.S.C.
// SECTION 108(a)(3)).

   var sWayBackCGI ="http://web.archive.org/web/20050308085053/";

   function xLateUrl(aCollection, sProp)
(     )var i = 0
     for(i = 0; i < aCollection.length; i++)
       if (aCollection[i][sProp].indexOf("mailto:") == –1 &&
          aCollection[i][sProp].indexOf("javascript:") == –1)
         aCollection[i][sProp] = sWayBackCGI + aCollection[i][sProp];
  }var i = 0
     for(i = 0; i < aCollection.length; i++)
       if (aCollection[i][sProp].indexOf("mailto:") == –1 &&
          aCollection[i][sProp].indexOf("javascript:") == –1)
         aCollection[i][sProp] = sWayBackCGI + aCollection[i][sProp];
  }var i = 0
     for(i = 0; i < aCollection.length; i++)
       if (aCollection[i][sProp].indexOf("mailto:") == –1 &&
          aCollection[i][sProp].indexOf("javascript:") == –1)
         aCollection[i][sProp] = sWayBackCGI + aCollection[i][sProp];
  }
  if (document.links)  xLateUrl(document.links, "href");
```

```
    if (document.images) xLateUrl(document.images, "src");
    if (document.embeds) xLateUrl(document.embeds, "src");

    if (document.body && document.body.background)
      document.body.background = sWayBackCGI +
document.body.background;

  //-->

  </SCRIPT>
  </html>
```

The problem with this method is that some links (embedded in scripts) will not be interpreted and therefore will continue to point to the original website. In some cases, interpretation of the page code activate some behavior, like a re-direction, even before this appended code is interpreted as modern browser do not wait to get the full document to interpret and display it.

Using a proxy that redirect all requests from the user's browser to the archive is the most efficient as mapping occurs after link interpretation is done by the interaction of the user (clicking) and the browser that interprets the code (HTML, client-side script, other formats) to generate the appropriate request. This is the most efficient as capacity of main browser to interpret code sets the standard for what is usually used on the Web. This approach requires setting up a proxy on that redirect to the archive, and to parameter a browser to use it, which can be too demanding for an online open archive environment. Use of browser plug-in to manage transition form the open to the proxy environment could alleviate this for end users.

Preferred Use

This method is appropriate for middle and large-scale archiving as well as for smaller archives that are concerned with preservation of content authenticity. As these methods store responses from the original server as it arrives to the client, without any transformation, it actually provides more faithfulness than the other methods. As it does not depend on any local file organization, it is also appropriate for migration as well as delivery of content.

Tools

This method requires an access infrastructure (see Chap. 6) as well as an archiving crawler (like Heritrix) and an index system for WARC files. The IIPC has developed a complete chain of tools to provide these functionalities.[22]

[22] See http://netpreserve.org/software/downloads.php (last visited May 2006).

1.4.3.3 Non-Web Archives

Description

In this approach illustrated in Fig. 1.7, documents that were on the Web are extracted from the hypertext context and re-organized in a different style in terms of access logic and/or format.

This can be the case when a set of documents taken from the Web is re-organized from a link-based access logic to a catalog-based one.

This is also the case when a page or even an entire website is transformed into PDF format. Adobe's Acrobat has this functionality (since version 6) and can transform an entire website into a single PDF document. In this case, the document is virtually printed which implies a frozen rendering and a paper page-like organization, even if linking can still work using an internal and proprietary naming scheme.

Comment

This approach makes sense mainly for objects that have originally been created and organized independently from the Web. This is the case for instance of large collections of digitized books, papers, music, videos made

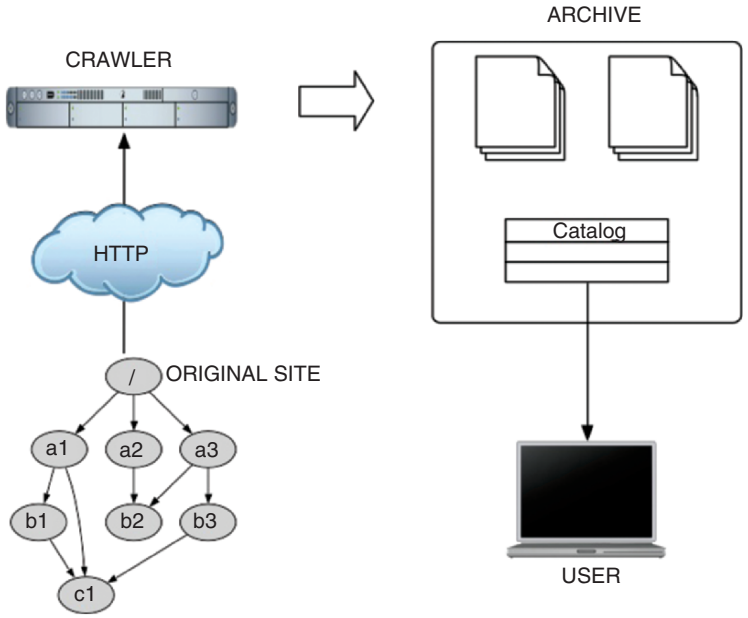


Fig. 1.7. Documents from the original site are re-organized in the archive, following a non-Web structure, using for instance a catalogue providing access to individual documents

available on the Web but which original organization was not hypertextual but catalog-based. It can be preferable in this case to stick to the original information architecture and archive these collections together with their catalogs merged in the archive's catalog. This is assuming that the hypertext context is deemed to be nonrelevant and can therefore be dismissed.

This has been the case for instance in the e-depot project of the KB in the Netherlands where scientific publications of Elsevier have been archived in a catalog-based system. The fact that Elsevier provides a Web access to this material has been considered as secondary to the content itself, structured as traditional scientific publication.

Preferred

This method is indicated for collections of content not structured in a Web-manner.

1.4.3.4 Summary

Table 1.2 summarize the various types of Web archives, their preferred used, tools, advantages and disadvantages.

Table 1.2. Summary of the Web archives types

Archive type	Local file system	Web-served	Non-Web
Description	All links are converted into relative ones. Hypertext Navigation is done directly on the local file system	A Web server is set up for access through which documents are served. Hypertext navigation is closed to the original one	Documents are extracted from the original hypertext content and re-organized along a different logic
Preferred use	Single site archiving and small and middle scale archiving	Small and middle scale archiving	Specific (non-Web) collections archiving
Tools	Website copier (like HTTrack)	Archiving Crawler (like Heritrix) and index system for WARC files	Depends on the final structuring of content

Advantages	Simple to implement	Authenticity, scalability	Enable integration in traditional catalogues or other local information architectures
Disadvantages	Does not scale up. Requires renaming and limited re-organization of content for hypertext navigation. Need a file system level management of archived collection and versions of items	Difficult to implement in absence of integrated software (this might change in the future)	Lost of hypertext structure. Can only be applied for isolated, non-Web documents

1.4.4 Quality and Completeness

Quality in general can be defined in a functional sense (fitting to particular use) or in an objective sense (matching to measurable characteristics). The term quality is applied to cultural collection in various contexts and senses. One can use it to qualify the state of conservation, the completeness of items or of the collection, the intellectual content level, etc. In each case, it relates to an ideal scale of perfection in a specific area (physical preservation, coverage of a domain, selection accuracy).

For Web archives, as we have seen, most imperfections come from the difficulty to gather content through the HTTP interface (see earlier section on "Client-side archiving") and the difficulty to render in a coherent manner the resulting content (see section on "Organization and storage"). Web archive's quality will therefore be mainly considered in this chapter as 1/the completeness of material (linked files) archived within a target perimeter and 2/the ability to render the original form of the site, particularly regarding navigation and interaction with the user (Masanès 2005).

Graphically, completeness can be measured horizontally by the number of relevant entry points found within the designated perimeter and vertically by the number of relevant linked nodes found from this entry point. Usually, entry points are site home pages, and links can direct the user either to

a new entry point (another site) or to elements of the same site. This is the case for site-oriented archiving.

In some cases, however, vertical inclusion is limited to embedded elements (images associated with a page for instance), and the collection is just organized horizontally, ignoring the site level. This is the case, for instance, for pure topic crawling where pages are not included based on their belonging to the site but only on their relevance to the topic.

Ideally, Web archives should be complete vertically as well as horizontally. But this is practically hard to achieve and priorities have to be set. Archiving is called "extensive" when horizontal completeness is preferred to vertical completeness (see Fig. 1.8).

This is the case, for instance, for the Internet Archive collection, which is donated by Alexa (as Burner 1997; Kimpton et al. 2006) explain, Alexa's crawler uses a breadth-first approach and adapts depth of crawl for a site according to traffic measured for this site).

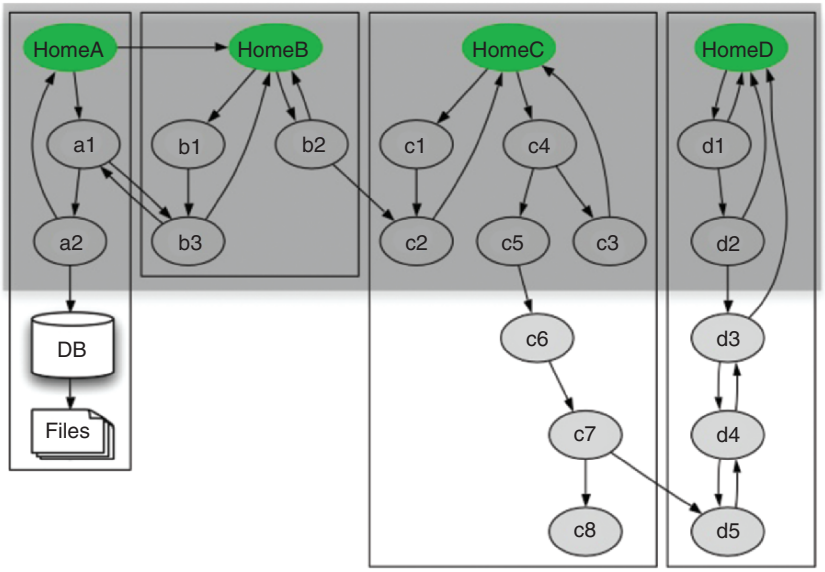


Fig. 1.8. Extensive collections, included more sites but archived at the surface level only. Only content in the shaded area will be archived. Pages deep in the hierarchy (c6, c7, c8, d3, d3, d5) as well as contend hidden behind database (hidden Web) will not be captured

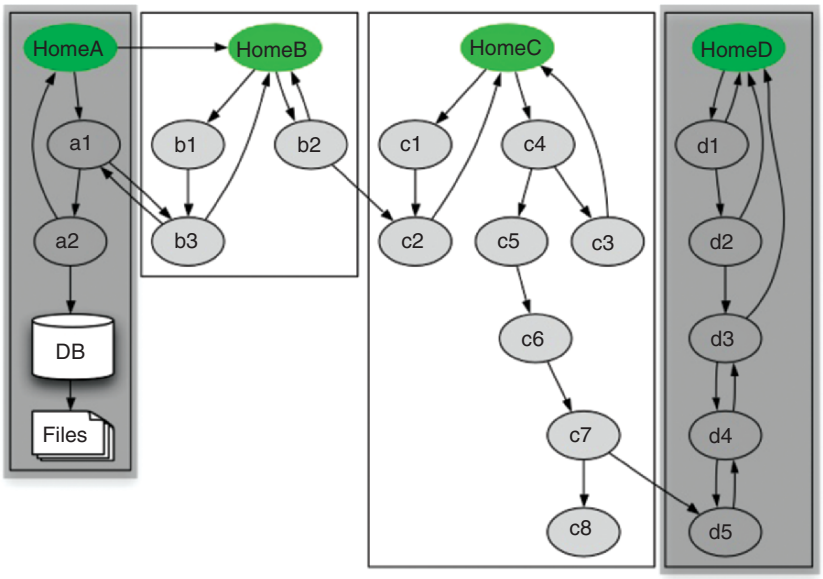


Fig. 1.9. Intensive archiving: less sites are crawled but crawl is done in depth. Only site A and D will be archived, but in totality, including the hidden Web portion of site A

Conversely, archiving is called "intensive" when vertical completeness is preferred to horizontal completeness (see Fig. 1.9).

This is the case, for instance, when a site-first priority is used for crawlers or when a manual verification is done with, where required, supplementary archiving. Intensive archiving is even more demanding for hidden Web sites (also called "Deep Web sites") where access to the full content is not possible with crawlers (for hidden Web archiving methodologies see Chap. 5).

1.5 Current Initiatives Overview

Web archives can be classified in several ways. We will review in this section the main ones, taking this opportunity to present some of the main initiatives in Web archiving and compare various approaches.

1.5.1 Archiving Actors

The type of organization creating and hosting the archive is the first criteria for classifying Web archives (WA). Some provide public access to their

collections (public WA), some do not (private WA). Amongst public WA, some provide online access; some provide on-site access in readings rooms (online public WA and offline public WA). Some also (and most of the time, primarily) manage nondigital collection (hybrid WA). Finally some are state-funded or nonprofit (noncommercial WA), whereas some are commercial companies (commercial WA).

Traditional heritage institutions (libraries, archives, museum), which are expanding their collections to the Web, form together the most part of the category of public hybrid WA. National libraries of several countries belong to this category (Sweden and Australia where the first back in 1996 and now there are many others).[23] National and regional and city archives are also starting archiving governmental and local authorities' websites.[24] Organization working on new forms of art like V2_ based in Rotterdam, Netherlands, are integrating net art in a general reflection and practice on preservation of unstable media (Fauconnier and Frommé 2004). All these archives can be categorized as noncommercial hybrid public WA as they integrate Web content in a larger context of collections. Most of them only provide offline access to documents at the moment.

Among them, the Library of Alexandria, Egypt, is one of the few providing online access to it's Web archive collection (mirroring the Internet Archive) and is an example of an online noncommercial hybrid public WA.

The pervasiveness of the Internet has also permitted the emergence of some new type of archiving organization holding only digital collection and providing access online, that we will categorize as public noncommercial online WA. The Internet Archive is the main example in this category (see Chapter 9 (Kimpton et al. 2006)).

Some commercial companies are archiving large collections of the public Web content as well, like Google with its "cache"[25] and Hanzo Archive for instance. These are example of online public commercial WA.

---

[23] Following them, several national libraries have started web archiving and have running programs (this list is not exhaustive): In Europe, Finland, Denmark, Norway, Iceland, France, Czech Republic, Slovenia, Italy, and Greece, in Asia Japan, China, and Singapore, the Library of Congress in the USA.

[24] The national archives of Australia (National Archives of Australia 2001), UK (Brown 2006), Canada, USA (Carlin 2004) have started systematic web archiving. See also the city of Antwerp DAVID's project (Boudrez and Eynde 2002).

[25] We do not consider here purely technical caching systems that contain copies of most of the content of the Internet, but in a very transient way (for a taxonomy of these systems see Dikaiakos (2004). On caching strategy and mechanism see, Krishnamurthy and Rexford (2001) and Hofmann and Beaumont (2005).

Finally, many organizations are developing internal Web archiving for their own purpose that we will classify as private WA. Qualifying the type of access (online or not) as well as their commercial status is less relevant here as these archives are only for private use.

1.5.2 Scope

Another useful way to classify Web archives is by considering the scope they adopt. Web archives can either be site-, topic-, or domain-centric.

1.5.2.1 Site-Centric Archiving

This type of archive, focused on a specific site, is mostly done by and for the creator of the site. This scoping is therefore mostly used for private WA. More and more companies for instance, being liable for all the content they publish, have to make sure they can refer back to older versions of their sites, blogs etc. This type of archives preferably uses site copiers (see chapter on the art of copying websites) and some services providers are emerging for this type of tailored internal archiving.[26]

1.5.2.2 Topic-Centric Archiving

Topic Web archiving is becoming more and more popular, often driven by direct research needs. While working on a specific field and its reflection on the Web, many scholars have confronted the ephemeral nature of Web publication, where the lifespan of Web sites is inappropriate for scientific verification (falsification requires access to the same data) as well as for long-lasting referral.[27]

This is the reason why several projects, often hosted in university libraries, have been undertaken to preserve primary material for research, such as the Digital Archive for Chinese Studies (DACHS) at Heidelberg University in Germany (see Chap. 10, Lecher 2006), or Archipol for analysis of Dutch political sites at Groningen University in the Netherlands, Voerman et al. 2002). These projects share not only a topic orientation but also the use of a network of informants (Lecher 2004, Lecher, 2006), that is, researchers who provide accurate and updated feeds for the archive.

---

[26] See hanzoarchives.com for instance.

[27] For use of web archives in the context of research see Chap. 2, "Web Use and Web Studies" and Chap. 10 "Academic archiving: DACHS'. See also Thelwall and Vaughan (2004) for a discussion of bias of web archives.

Other topic-centric projects have been carried on in libraries by actively seeking and archiving electoral Web sites, such as the Minerva project from the Library of Congress (Schneider et al. 2003) or the French elections Web archive made by the Bibliothèque nationale de France (Masanès 2005). Compared to the previous topic-centric approach, discovery of sites does not come naturally as a by-product of research activity and needs to be undertaken as a specific activity.[28]

Finally some project pertaining to this category use topic crawling for discovery and capture of content related to the same topic (Chakrabarti et al. 1999; Bergmark 2002; Bergmark et al. 2002; Qin et al. 2004), see also Chap. 5 (Masanès 2006b). Automatic discovery and filtering is done using traditional crawling technique combined with a page level appraisal of textual content sometime blended with some link structure mining. The proximity with the topic can be "learned" from a corpus or from user feed-back. Although promising, this area still requires research to be applied for archiving.

1.5.2.3 Domain-Centric Archiving

Archive building can also be done based on location of content. This characterizes a third type WA. "Domain" is used here in the network sense of the word or, by extension, in the national sense of the term, which is a combination criteria for targeting sites of a specific country.[29]

The DNS allows a simple and actionable selection of content based on domain names. The fact is that domain names, even for the upper levels domains managed by official delegation from the ICANN, do not really follow rules with regards to naming, functional specialization and organization, but rather traditions (Liu and Albitz 1999), see also on the evolution on Internet naming (Koehler 1999). However, one can distinguish functional or generic types (like .com and .edu) and geographical types (.ch and .jp)[30] types for the first level domain (often called Top-Level domain). The geographical top-level domains often have functional subdivisions (like asso.fr, gob.mex), which means that the second-level domain (SLD) will also be managed in the same way. There are some exceptions to the tradition like for the .us

[28] On selection see Chap. 3 (Masanès 2006b).

[29] For a discussion of the possible way of delimiting a national Internet space see Arvidson et al. (2000), Abiteboul et al. (2002), Lampos et al. (2004). For studies of national internet space characteristics see, Baeza-Yates et al. (2005a, 2005b), and Gomes and Silva (2003).

[30] This follows the ISO 3166 two-letters country names standard, except for .uk which should be .gb, and except also that, it is currently extended to three letters for regions, like with the .cat domain for Catalonia in Spain.

TLD which has other geographical subdivision (by states). Note, however, that all these portions of the Internet domain space being managed by delegation,[31] each entity in charge of them can apply a specific policy regarding allocation and control of their space, therefore making utilization of TLD or SLD for Web archiving selection dependent on each case on assessment of this policy (the .org and .com gTLD for instance are used by all type of organizations and not only by commercial ones for the .com and nonprofit ones for the .org, as there are no restrictions for registration). In addition to this, some entities in charge of TLD's management change their policy with time (.org and .net used to have restrictions before 1996 and .fr TLD has significantly reduce restrictions in 2005 for instance).

This being said, let us recall the great advantage that brings criteria that can be automatically detected by crawlers, like domain names. Several projects actually implement the domain-centric approach. Some focus on a generic domain like .gov (Cruse et al. 2003; Carlin 2004) or .edu (Lyle 2004). Some use national domain, like Kulturarw started in 1997 by the Swedish Royal Library (Arvidson et al. 2000), which covers the .se TLD and also Swedish pages linked from it and located in generic domains such as .com.

1.5.3 Methods Used

Projects can also noticeably differ with respect to the methodological approach they take for discovery, acquisition, and description of content. An important difference that spans across all these phases is the use of manual versus automatic processing. Although the apparent simplicity of this opposition has to be balanced as automatic processing occurs at several levels (capture, use of search engines for "manual discovery", etc. (Masanès 2006b)), it remains that WA can be categorized according to this opposition, which impacts directly on scalability and quality of archives.

As can be anticipated, automation of these tasks enables a tremendous lowering of the cost per site archived.[32] Ideally, a single operator running a

[31] On DNS governance and its political implications see Mueller (2002).

[32] Phillips (2005) provides very useful detailed time and costs estimations of manual processing of sites for one of the most ancient existing web archive. Time estimates are the following (excerpt):

- Identification and selection: 30 min;
- Gathering, quality assurance, and archiving instances: 210 min;
- Cataloguing: 81 min;

We lack similar precise estimates for automatic discovery, capture, and indexing (instead of cataloging), but it is, at the exception of quality assurance, several order of magnitude below.

crawl can "discover" and download millions of pages. Considering that full-text indexing provides a powerful finding aid comparable if not superior to cataloging in many cases, then we can see that, here again, how automation lowers costs dramatically, as it can be applied on a large scale (Stack 2005), see also Chap. 6 (Hallgrímsson 2006).

Unfortunately, automation reaches some limits, and manual handling must be done in some cases. Discovery, for instance, can be done manually or automatically. When done manually, it can be a specific activity or a by-product of other activities, as the DACHS (Lecher 2006) and Achipol (Voerman et al. 2002) Web archives show. This type of approach is usually taken for topic-centric archiving. Although topic crawling has proven efficiency for the discovery of topic-related sites or pages, automatic tools can certainly not yet compare with a network of experts providing references to the best material they are aware of.

However, a lack of domain expertise and understanding is not the only disadvantage crawlers have. Also to be considered is the delay needed to find new sites. It can take lots of time for holistic crawl to discover sites. When it comes to ephemeral sites, related to an event for instance, the delay can be too long to locate and archive them. This difference has been studied by (Masanès 2005) with a comparison of sites discovered by Alexa's crawler and accessible today on the Internet Archive and sites related to the French elections of 2002 located by a team of reference librarians and archived by the national library of France. This study shows a clear advantage to manual active selection in event-related collections for timely discovery and in-depth focus.

Classifying WA according to their methodology could also be done at a finer grain. Beyond the dichotomy manual/automatic processing, one could consider for instance the type or source used for discovery, the periodicity of search and capture, the level of quality verification made, the granularity of archived items (sites, pages), etc.

It is a fact however that most of WA tend to fit in two main models, the main differentiator being whether selection is done manually or not. One is the model of holistic crawls, usually domain-centric (national domains or generic domains) or open (Internet Archive), the other is the model of individual selection of a limited number of seeds or entry points (usually sites) done manually. Finer distinctions in their methodological approach are rarely noticed nor used to classify them.

1.6 Conclusion

The Web has only fifteen years of existence and one could say, that conservation of its memory has started relatively early compared to other media.[33] But we have only made the first necessary steps for its preservation. Current state of preservation relies on too few institutions and does not achieve so much coverage. Roles and the responsibilities are far from being clear to most stakeholders, and sustainability of many of the most significant collections is not granted. And we are still in a period where no technological rupture has taken place since the Web's inception. Current browsers together with a limited number of plug-ins can handle most of the formats that can be found on the Web (see Chap. 8 for a detail overview of preservation of Web material, Day 2006). But this situation will not last for ever and Web preservation will encounter a serious challenge when major technological change occurs on the Web (which may not be called like this afterwards).

It is thus encouraging to see that more and more heritage institutions are engaging in Web archiving. A recent survey by the Research Library Group (RLG 2006) showed that 60% of their members considered that Web archiving was part of their mission (RLG 2006), which is very heartening. We hope that the presentation made in this chapter of the main issues and methods together with their rationale will help them and others to participate in this collective effort.

References

Aarseth, E. J. (1997). *Cybertext: perspectives on ergodic literature.* Baltimore, MD: Johns Hopkins University Press

Abiteboul, S., Cobena, G., Masanès, J., & Sedrati, G. (2002). *A first experience in archiving the French Web*. Paper presented at the Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries

Abiteboul, S., Preda, M., & Cobena, G. (2003). *Adaptive on-line page importance computation*. Paper presented at the Proceedings of the twelfth international conference on World Wide Web

Antoniol, G., Canfora, G., Cimitile, A., & De Lucia, A. (1999). *Websites: files, programs or database*. Paper presented at the 1st International Workshop on Web Site Evolution, Atlanta, USA

Arvidson, A., Persson, K., & Mannerheim, J. (2000). *The Kulturarw3 project - The Royal Swedish Web Archiw3e – An example of "complete" collection of*

[33] Most of radio and TV broadcasts in the world are still not preserved at all.

web pages. Paper presented at the 66th IFLA – International Federation of Library Associations and Institutions, Jerusalem

Baeza-Yates, R. & Castillo, C. (2005). Characteristics of the Web of Spain. *Cybermetrics*, *9*

Baeza-Yates, R., Castillo, C., & Efthimiadis, E. (2005a). Characterization of national Web domains

Baeza-Yates, R. A., Castillo, C., Marin, M., & Rodriguez, A. (2005b). *Crawling a country: better strategies than breadth-first for Web page ordering*. Paper presented at the WWW 05: Proceedings of the 14th international conference on World Wide Web, Chiba, Japan

Balayé, S. (1988). *La Bibliothèque nationale, des origines à 1800* (Histoire des idées et critique littéraire; vol. 262). Genève: Droz

Battelle, J. (2005). Google Announces New Index Size, Shifts Focus from Counting. http://battellemedia.com/archives/001889.php

Benjamin, W. (1963). *Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit; drei Studien zur Kunstsoziologie.* [Frankfurt am Main]: Suhrkamp

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

Bergmark, D. (2002). *Collection synthesis*. Paper presented at the 2nd ACM/ IEEE-CS joint conference on Digital libraries, Portland, USA

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002). *Focused crawls, tunneling, and digital libraries*. Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries, Roma, Italy

Berners-Lee, T. & Connolly, D. (1995). Hypertext Markup Language – 2.0. *RFC*, *1866*

Berners-Lee, T. (1994). Universal Resource Identifiers in WWW, A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web. *RFC 1630*

Berners-Lee, T. (1998). Cool URIs don't change. http://www.w3.org/Provider/ Style/URI.html

Berners-Lee, T. & Fischetti, M. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor* (1st pbk. ed.). New York: HarperCollins

Björneborn, L. & Ingwersen, P. (2001). Perspective of webometrics. *Scientometrics*, *50*(1), 65–82

Bolter, J. D. (2001). *Writing space: Computers, hypertext, and the remediation of print* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates

Borgman, C. L. (2000). *From Gutenberg to the global information infrastructure: access to information in the networked world* (Digital libraries and electronic publishing). Cambridge, MA: MIT

Borgman, C. L. (2003). The Invisible Library: Paradox of the Global Information Infrastructure. *Library Trends*, *51*(4), 652–674

Boudrez, P. & Eynde, V. D., Sofie. (2002). Archiving Websites

Boufkhad, Y. & Viennot, L. (2003). The Observable Web. *RR*

Boyko, A. (2004). Test Bed Taxonomy. *IIPC Reports,* 16

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., & Stata, R., et al. (2000). *Graph structure in the web*. Paper presented at the 9th International World Wide Web Conference (WWW9), Amsterdam, Netherlands

Brown, A. (2006). *Archiving the Web: A guide for information management professionals*. Library Assn Pub.

Brügger, N. (2005). *Archiving Websites, general considerations and strategies*. Aarhus, Denmark: Center for Internet Research

Bruns, A. (2005). *Gatewatching: Collaborative online news production* (Digital formations, v. 26). New York: P. Lang

Burner, M. (1997). Crawling towards Eternity Building An Archive of The World Wide Web. *New Architect*, *5*

Canfora, L. (1989). *The vanished library* (Hellenistic Culture and Society; 7). Berkeley: University of California Press

Canfora, L. (1996). Les bibliothèques anciennes et l'histoire des textes. In M. Baratin, & C. Jacob (Eds.), *Le pouvoir des bibliothèques: la mémoire des livres en Occident*. (pp. 338 p). Paris: A. Michel

Carlin, J. W. (2004). Harvest of agency public websites. *NARA Bulletin*, *2005-02*

Castells, M. (1996). *The rise of the network society*. Malden, MA: Blackwell

Castillo, C., Marin, M., Rodriguez, A., & Baeza-Yates, R. A. (2004). Scheduling Algorithms for Web Crawling

Chakrabarti, S. (2002). *Mining the Web: discovering knowledge from hypertext data*. San Francisco, CA: Morgan Kaufmann

Chakrabarti, S., Berg, M. V. D., & Dom, B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, *31*, 1623–1640

Chang, K. C.-C., He, B., Li, C., Patel, M., & Zhang, Z. (2004). Structured databases on the web: observations and implications. *SIGMOD Record*, *33*(3), 61–70

Charlesworth, A. (2003). *Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia*

Cho, J., & Garcia-Molina, H. (2000). *The evolution of the web and implications for an Incremental Crawler*. Paper presented at the Proceedings of the 26th International Conference on Very Large Data Bases

Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient Crawling Through url ordering. *Computer Networks and Isdn Systems*, *30*, 161–172

Christensen-Dalsgaard, B. (2001). *Archive experience, not data*. Paper presented at the Preserving the Present for the Future - Strategies for the Internet, The Royal Library, Copenhagen, Denmark

Crowston, K., & Williams, M. (1997). *Reproduced and emergent genres of communication on the World-Wide Web*. Paper presented at the 30th Annual Hawaii International Conference on System Sciences (HICSS-30), Wailea, USA

Cruse, P., Eckman, C., & Kunze, J. (2003). Web-based government information: Evaluating solutions for capture, curation, and preservation. *An Andrew W. Mellon funded initiative of the California Digital Library*

Dahn, M. (2000). Counting Angels on a Pinhead: Critically Interpreting Web Size Estimates. *Online*, *January/February*, 35–40

Day, M. (2006). The long-term preservation of Web content. In J. Masanès (Ed.), *Web archiving*. Berlin Heidelberg New York: Springer

Dikaiakos, M. D. (2004). Intermediary infrastructures for the World Wide web. *Computer Networks*, *45*(4), 421–47

Dobra, A., & Fienberg, S. E. (2004). How Large Is the WorldWide Web?. In M. Levene, & A. Poulovassilis (Eds.), *Web dynamics web dynamics – adapting to change in content, size, topology and use*. (pp. 23–44). Berlin Heidelberg New York: Springer

Dubberly, H., Forlizzi, J., Hodge, C., Laurel, B., Lyman, P., Meggs, P. B., et al. (2002). Archiving experience design, a virtual roundtable discussion. *LOOP: AIGA Journal of Interaction Design Education*, *Number 6*

Dumais, S. T., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R., et al. (2003). *Stuff I've seen: A system for personal information retrieval and re-use*. Toronto, Canada

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. *Journal of Information Science*, *26*(5), 329–335

Eiron, N. & McCurley, K. S. (2003). *Locality, hierarchy, and bidirectionality on the Web*. Paper presented at the Workshop on Web Algorithms and Models

Eisenstein, E. L. (1979). *The printing press as an agent of change: Communications and cultural transformations in early modern Europe.* Cambridge [Eng.]; New York: Cambridge University Press

Entlich, R. (2004). Blog Today, Gone Tomorrow? Preservation of Weblogs. *RLG DigiNews*, *8*(4)

Eriksen, L. B. & Ihlström, C. (2000). *Evolution of the web news genre – The slow move beyond the print metaphor*. Paper presented at the 33rd Hawaii International Conference on System Sciences (HICSS-33), Hawaii, USA

Estivals, R. (1961). *Le dépôt légal sous l'Ancien Régime, de 1537 à 1791*. Paris: M. Rivière

Estivals, R. (1965). *La statistique bibliographique de la France sous la monarchie au XVIIIe siècle.* Paris: Mouton

Fauconnier, S. & Frommé, R. (2004). Capturing unstable media, summary of research

Fayet-Scribe, S. (2000). *Histoire de la documentation en France: Culture, science, et technologie de l'information,* 1895–1937 (CNRS histoire). Paris: CNRS

Featherstone, M. (2000). Archiving cultures. *British Journal of Sociology*, *51*(1)

Febvre, L. P. V. & Martin, H. J. (1976). *The coming of the book: The impact of printing* 1450–1800 ([New ed.] ed.). London: NLB

Fetterly, D., Manasse, M., Najork, M. & Wiener, J. (2003). *A large-scale study of the evolution of web pages*. Budapest, Hungary

Fielding, R. T., Gettys, J., Mogul, J., Nielsen, H. F., Masinter, L., J, P., et al. (1999). Hypertext Transfer Protocol – HTTP/1.1. *RFC*, *2616*

Fitch, K. (2003). *Web site archiving: An approach to recording every materially different response produced by a website*. Paper presented at the AusWeb

2003: The Ninth Australian World Wide Web Conference, Sanctuary Cove, Australia

Florescu, D., Levy, A., & Mendelzon, A. (1998). Database techniques for the World-Wide Web: A survey. *SIGMOD Record 27*, 59–74

Freeman, E. & Gelernter, D. (1996). Lifestreams: A storage model for personal data. *SIGMOD Record*, *25*(1), 80–86

Gemmell, J., Bell, G., Lueder, R., Drucker, S., & Wong, C. (2002). *MyLifeBits: fulfilling the Memex vision*. Juan-les-Pins, France

Gibson, D., Punera, K., & Tomkins, A. (2005). *The volume and evolution of web page templates*. Paper presented at the WWW'05 14th international conference on World Wide Web, Chiba, Japan

Gillies, J. & Cailliau, R. (2000). *How the Web was born: The story of the World Wide Web.* Oxford: Oxford University Press

Golder, S. & Huberman, B. A. (2005). The Structure of Collaborative Tagging Systems

Gomes, D. & Silva, M. J. (2003). *A Characterization of the Portuguese Web*. Paper presented at the 3rd Workshop on Web Archives (IWAW'03), Trondheim, Norway

Gulli, A. & Signorini, A. (2005). *The indexable web is more than 11.5 billion pages*. Chiba, Japan

Halavais, A. (2004). Tracking Ideas in the Blogosphere

Hallgrímsson, T. (2006). Access and finding aids or web archives. In J. Masanès (Ed.), *Web archiving.* Berlin Heidelberg New York: Springer

Hine, C. (2000). *Virtual ethnography.* London; Thousand Oaks, CA: Sage

Hofmann, M. & Beaumont, L. R. (2005). *Content networking: Architecture, protocols, and practice* (The Morgan Kaufmann Series in Networking). Amsterdam; Boston: Morgan Kaufmann

Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, *54*(2)

Jones, S. & Johnson, C. (2006). Web Use and Web Studies. In J. Masanès (Ed.), *Web Archiving.* Berlin Heidelberg New York: Springer

Jones, W., Bruce, H., & Dumais, S. (2001). *Keeping found things found on the web*. Atlanta, GA, USA

Jones, W., Bruce, H., & Dumais, S. (2003). *How do people get back to information on the Web? How can they do it better?* Paper presented at the IFIP INTERACT'03

Kahle, B. (1997). Preserving the Internet. *Scientific American*, *397*, 82–84

Kahle, B. (2002). The Internet Archive. *RLG DigiNews*, *6*(3)

Kimpton, M., Braggs, M., & Ubois, J. (2006). Year by Year: From an Archive of the Internet to an Archive on the Internet. In J. Masanès (Ed.), *Web Archiving.* Berlin Heidelberg New York: Springer

Koehler, W. (1999). Unraveling the ISSUES, ACTORS, & ALPHABET SOUP of the Great Domain Name Debates. S*earcher*, *7*(5)

Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research*, *9*(2)

Krishnamurthy, B. & Rexford, J. (2001). *Web protocols and practice: HTTP/1.1, networking protocols, caching, and traffic measurement.* Boston, MA: Addison-Wesley

Lagoze, C., Dean B. K., Sandy, P., & Jesurogaii, S. (2005). What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. *D-Lib Magazine*, 11–11

Lampos, C., Eirinaki, M., Jevtuchova, D., & Vazirgiannis, M. (2004). *Archiving the Greek Web*. Paper presented at the 4th International Web Archiving Workshop (IWAW'04), Bath (UK)

Landow, G. P. (1997). *Hypertext 2.0* (Rev., amplified ed.). Baltimore: Johns Hopkins University Press

Lavoie, B. F. & Schonfeld, R. C. (2005). *The systemwide print book collection*. Paper presented at the CNI Spring 2005 Task Force Meeting

Lawrence, S. & Giles, C. L. (1998). Searching the Web. *Science*, *281*, 175.

Lawrence, S. & Giles, C. L. (1999). Accessibility of Information on the Web. *Nature*, *400*, 107–109

Lecher, H. E. (2004). *Informant networks, alarm systems, and research contributors. Selection and ingest process for the Digital Archive for Chinese Studies*. Paper presented at the Archiving Web Resources Conference – Issues for Cultural Heritage Institutions, NLA, Canberra, Australia

Lecher, H. E. (2006). Academic Web archiving: DACHS. In J. Masanès (Ed.), *Web archiving.* Berlin Heidelberg New York: Springer

Levy, P. (1997). *Collective intelligence: Mankind's emerging world in cyberspace.* Cambridge, MA: Perseus Books

Liu, C. & Albitz, P. (1999). *DNS & BIND* (3rd ed.). O'Reilly & Associates

Lueg, C. & Fisher, D. (2003). *From Usenet to CoWebs: Interacting with social information spaces* (Computer supported cooperative work). Berlin Heidelberg London New York: Springer

Lyle, J. A. (2004). *Sampling the Umich.edu Domain*. Paper presented at the 4th International Web Archiving Workshop (IWAW'04), Bath (UK)

Lyman, P. (2002). Archiving the World Wide Web. In CLIR (Ed.), *Building a national strategy for preservation: issues in digital media archiving.* Council on Library and Information Resources and the Library of Congress

Lyman, P. & Kahle, B. (1998). Archiving digital cultural artifacts. *D-Lib Magazine*

Mantratzis, C. & Orgun, M. (2004). *Towards a peer2peer world-wide-web for the broadband-enabled user community*

Masanès, J. (2002). Towards continuous Web archiving: First results and an agenda for the future. *D-Lib Magazine*, *8*(12)

Masanès, J. (2004). Site-first priority: Implementing the frontline

Masanès, J. (2005). Web archiving methods and approaches: A comparative study. *Library Trends*

Masanès, J. (2006a). Collecting the hidden web. In J. Masanès (Ed.), *Web archiving.* Berlin Heidelberg New York: Springer

Masanès, J. (2006b). Selection for Web Archives. In J. Masanès (Ed.), *Web archiving.* Berlin Heidelberg New York: Springer

Mohr, G., Kimpton, M., Stack, M. & Ranitovic, I. (2004). *Introduction to Heritrix, an archival quality web crawler*. Paper presented at the 4th International Web Archiving Workshop (IWAW'04), Bath (UK)

Mueller, M. (2002). *Ruling the root: Internet governance and the taming of cyberspace.* Cambridge, MA: MIT

Najork, M. & Heydon, A. (2001). High-performance Web crawling. *SRC Research Report*

Najork, M. & Wiener, J. (2001). *Breadth-first search crawling yields high-quality pages*. Paper presented at the 10th World Wide Web Conference (WWW'10), Hong Kong

National Archives of Australia. (2001). Archiving Web resources: A policy for keeping records of web-based activity in the Commonwealth Government

Osborn, T. (1999). The ordinariness of the archive. *History of the human sciences*, *12*(2)

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The Pagerank citation ranking: Bringing order to the Web, 17

Pandey, S. & Olston, C. (2005). *User-centric Web crawling*. Chiba, Japan

Pant, G., Srinivasan, P. & Menczer, F. (2004). Crawling the Web. In M. Levene, & A. Poulovassilis (Eds.), *Web Dynamics*. (pp. 153–178). Berlin Heidelberg New York: Springer

Pastor-Satorras, R. & Vespignani, A. (2004). *Evolution and structure of the Internet: A statistical physics approach.* Cambridge, UK; New York: Cambridge University Press

Phillips, M. E. (2005). Selective archiving of Web Resources: A study of acquisition costs at the National Library of Australia. *RLG DigiNews*, *9*(3)

Qin, J., Zhou, Y. & Chau, M. (2004). *Building domain-specific web collections for scientific digital libraries: A meta-search enhanced focused crawling method*. Tuscon, AZ, USA

Rekimoto, J. (1999). *Time-machine computing: A time-centric approach for the information environment*. Paper presented at the 12th annual ACM symposium on User interface software and technology, Asheville, North Carolina, USA

Riché, P. (1996). La bibliothèque et la formation de la culture médiévale. In M. Baratin, & C. Jacob (Eds.), *Le pouvoir des bibliothèques: la mémoire des livres en Occident* (p. 338). Paris: A. Michel

Ringel, M., Cutrell, E., Dumais, S., Horvitz, E. (2003). *Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores*. Paper presented at the IFIP INTERACT '03

RLG. (2006). Web Archiving Program. http://www.rlg.org/en/page.php?Page_ID=399

Roche, X. (2006). Copying web sites. In J. Masanès (Ed.), *Web Archiving.* Berlin Heidelberg New York: Springer

Rosenfeld, L. & Morville, P. (2002). *Information architecture for the World Wide Web* (2nd ed.). Cambridge, MA: O'Reilly

Scharl, A. (2000). *Evolutionary Web development* (Applied computing). Berlin Heidelberg New York: Springer

Shepherd, M. & Polanyi, L. (2000). *Genre in Digital Documents*. Paper presented at the Proceedings of the 33rd Hawaii International Conference on System Sciences – vol. 3

Sonnenreich, W. (1997). A History of Search Engines. http://www.wiley.com/legacy/compbooks/sonnenreich/history.html

Spinellis, D. (2003). The decay and failures of web references. *Communications of ACM*, *46*(1), 71–77

Stack, M. (2005). *Full Text Search of Web Archive Collections*. Paper presented at the IWAW'05, Vienna, Austria

Star, S. L. & Ruhleder, K. (1994). *Steps towards an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems*. Chapel Hill, NC, United States

Teevan, J. (2004). How people re-find Information when the Web changes. AIM-2004-012

Thelwall, M. (2001). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, *52*(13), 1157–1168

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of American Society of Information Science and Technology 57*(1), 60–68

Thelwall, M. & Harries, G. (2004). Do the websites of higher rated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology*, *55*(2), 149–59

Thelwall, M. & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet archive. *Library & Information Science Research*, *26*(2), 162–176

Ubois, J. (2002). The Oakland archive policy. Recommendations for managing removal requests and preserving archival integrity

Voerman, G., Keyzer, A., Hollander, F. D., & Druiven, H. (2002). Archiving the Web: Political Party Web sites in the Netherlands. *European Political Science*, *2*(1)

2 Web Use and Web Studies

Steve Jones and Camille Johnson

University of Illinois at Chicago
sjones@uic.edu

2.1 Summary

In 2002, the Online Computer Library Center (OCLC) estimated that over three million publicly accessible websites existed on the World Wide Web (O'Neill et al. 2003, para. 9). In the material world of information, this number would approximate from 14 to 28 million books, matching or exceeding the number of volumes held by most of the world's largest libraries. This overwhelming collection of information represents an endless, accessible supply of both verbal and visual data for scholars interested in studying online activity. And although the sheer volume of available resources on the Web has presented the challenge of choosing *what* to study and *how* to study, the Web has proven even more daunting. In this chapter, we will provide an overview of methodological approaches researchers have used to study the Web.

The goal in doing so is not to provide an exhaustive categorization of methodologies. Instead, it is our hope that by understanding the methods used for studying the Web and studying Web use, those who seek to archive and preserve the Web can better understand the needs of the research community.

The Web consists of an immense variety of types of materials whose variety is best understood along two dimensions. First, the Web itself is a medium, and not simply content. More to the point, it is both a medium that conveys content via numerous protocols (such as HTTP), and it is also a "container" for content, one which further "shapes" content and also presents it to viewers. However, the presentation of content is at the viewer's discretion and is further shaped by the viewing tools used (browsers and other applications). In other words, though the content may be the same Web page, two viewers using different browsers or using different browser settings may ultimately see different pages. Second, and further unlike the

world of analog materials, the very definition of media for storing the Web is open to question. Websites can be locally stored or cached, may be mirrored, may be dynamic, or may be virtually ephemeral, as is the case with Web cams.

Internet research, in general, and Web research, in particular, is populated by a diverse group of scholars. Their various disciplines include linguistics, journalism, political science, business administration, geography, advertising, communication, and the arts, to name only a few. Because we are dealing with such a wide array of academic traditions, the types of materials with which they work are very diverse, and the working definitions of research methods vary as well. What counts as "ethnographic" analysis of Web texts for a marketing scholar may be interpreted as a qualitative content analysis by a communication scholar. Therefore, another goal of this chapter is not to provide rigid Web content categories or blueprints for each methodological approach as they apply to Web studies, but rather to present a range of interpretations and applications of these methods as they have proven most useful to researchers of the Web.

2.2 Content Analysis

Content analysis is one of the more common methods for studying the Web. As with other media, a researcher who uses content analysis codes Web content, either written text or images, based on particular criteria and places them within relevant categories or themes; in other words, it is a survey of Web content, rather than Web users. Within Web studies, content analysis has primarily been used as a comparative tool, allowing the researcher to make meaningful comparisons of content between similar Web texts. A study of antiglobalization organizations' websites used content analysis to determine whether cohesion in message and purpose existed among the sites (Van Aelst and Walgrave 2002). Content was coded and sorted based on what appeared to be its primary function within four areas: to provide information about the organization, to provide information about antiglobalization issues, to foster interactivity with the group organizers and other members, and to foster mobilization on behalf of the cause, such as donating money signing an online petition. Through their analysis, Van Aelst and Walgrave were able to conclude that indeed, antiglobalizations organizations online were generally using their websites in similar ways to inform and involve their memberships.

Another comparative study analyzed the content of radio station websites to determine how the commercial radio industry was responding to

the availability of the World Wide Web for promoting their stations. This assessment was made by analyzing the types of user-based information being provided on radio stations' websites (e.g., traffic maps and program logs), their use of websites as tools for promoting their station (e.g., contest information and DJ bios), and their incorporation of interactive features (e.g., e-mail addresses for the stations' staff members and listener surveys) (Pitts and Harms 2003). For radio stations, this Web study provided a valuable resource for evaluating their current and potential uses of the Web.

Content analysis has also been used to study the impact of institutional policy changes on related websites. Educational policies, for example, have placed increasing importance on the incorporation of information and communications technology (ICT) into secondary schools' institutional missions. One study chose to evaluate the progress being made toward achieving this goal as outlined by the National Grid for Learning in the UK through a content analysis of 150 secondary school websites (Hesketh and Selwyn 1999). Images and written texts were coded, identifying the websites within one of the following attitude categories: Proactive, pupil-centered, or reactive. Through these categories and an analysis of the schools' profiles, the researchers recognized a correlation between a commitment to ICT integration and institutional capital; the more economic and social capital a school had, the more likely that their website's content would reflect a positive, proactive attitude toward the incorporation of ICT in their mission.

The US Federal Trade Commission has employed content analysis to evaluate the implementation of online privacy and information security notices on commercial websites (Milne and Culnan 2002). Their study comprised a longitudinal analysis of four Web "surveys" spanning from 1998 to 2001. Criteria such as the inclusion or exclusion of notices regarding the disclosure of visitor information with third parties and use of "cookies," collection of user information beyond e-mail addresses, and the use of information security on the sites were used to determine whether the websites met the requirements of fair information practices. The results from the 2001 analysis were also compared with previous years and indicated that among commercial websites that collect personal information from their users there has been an increase in the implementation of fair information practices (p. 355).

2.3 Surveys

Web surveys can be defined in two distinct ways: those which are surveys *on* the Web, and those which are surveys *of* the Web. The former type appears to be the most common, with researchers posting surveys on websites in order to access their population of interest. In most cases, their purpose is to collect information about people's uses of the Internet,[1] such as the Graphic, Visualization, and Usability Center's (GVU) World Wide Web User Survey (1998). The GVU project is an ongoing study of Internet use that began in 1994. The project recruits participants through Internet-related newsgroup postings, banner advertisements on news media and search engine websites, and announcement in offline media such as magazines and newspapers. Several focused surveys have been generated by the GVU including a general demographics survey, a technology demographics survey which collected info such as connection speed of users and Internet browser used, a survey of users' attitudes toward online privacy and security, a general survey of computer, Web and Internet use, as well as a survey of online product searching and purchasing activities. Sample sizes typically topped 1,000, with the general demographics survey surpassing 5,000 participants.

Organizations have also found Web surveys a useful tool for assessing the quality of user experiences using related websites or Web functions. One such study conducted a survey of approximately 450 museum website visitors, collecting both demographic information and responses to the quality and usability of the sites (Sarraf 1999). Many museums have begun to provide websites as a means for supplementing information about their collections as well as creating an interactive, accessible relationship with potential and past visitors (p. 1). Qualitative and quantitative responses were solicited, providing rich, detailed feedback for the museums involved. Carswell and Venkatesh (2002) were able to use a Web-based survey to solicit evaluative responses from over 500 graduate students who

[1] With Internet access reaching 59% among US adults in 2002 (Spooner, 2003), researchers in the US are also beginning to consider Web-based surveys a practical substitute for traditional types of survey administration, such as paper and telephone, on topics other than Internet uses and attitudes. Researchers interested in collecting timely responses to current events seem particularly interested in Web-based surveys, exemplified by studies the impact of the terrorist attacks in New York City on September 11, 2001 on participants (e.g., Lee et al. 2003). Just nine days after the attacks, one team of researchers was able to distribute a survey via the Web analyzing the psychological responses of Americans to this traumatic event (Silver et al. 2002).

had participated in asynchronous online courses. They found partial support for their hypotheses that acceptance of and future intent to use technology in an asynchronous online course would be positively influenced by the student's attitudes and perception of the technology. In all cases, including the GVU project, the benefit of using a Web-based survey was clear: the populations of interest were Web-users, and the data collected related directly to their Web activities.

2.4 Rhetorical Analysis

Another subset of Web research looks critically at Web texts to identify persuasive strategies through rhetorical analysis. As defined by Warnick (1998)

> …a rhetorical critical method considers how the text gives presence to some elements as opposed to others, how narrative constructions configure browser experience in certain ways, and how the discourse plays to the predispositions and habits of mind of its audience. (p. 309)

In her study of political websites during the 1996 US presidential campaign, Warnick specifically analyzed the rhetorical strategies of parodic websites – sites which mimicked the design and content of legitimate campaign websites, possibly to erode the credibility of the sites mimicked and certainly to provide comic relief (p. 308). She concludes that the parodic sites relied on popular narratives of distrust in government, the simulation of political participation through interactive features such as petitions which were never delivered to their intended party, and innuendo regarding candidates' political and criminal pasts to persuade their readers to adopt their rhetorical vision. Ultimately, Warnick finds the strategies hypocritical, as they include many of the same unethical behaviors purportedly used by the candidates in question.

Rhetorical analysis of websites has been used to identify rhetorical communities, those groups which use the Web to create their own worldview which may or may not agree with popular narratives. Kroeber (2001) discusses how a community of websites created by feminist mothers presented a platform for participants to actively resist negative, feminist conceptualizations of motherhood. The women were able to renegotiate feminism through the written text and images found on their websites, arguing for an understanding of motherhood as an empowering, even integral, part of being a feminist.

Another study looked at the use of websites for persuasive purposes by hate groups such as the Knights of the Ku Klux Klan, Nation of Islam, and the Neo-Nazi organization National Alliance (Duffy 2003). In this case, Duffy used a rhetorical critical approach called Fantasy Theme analysis, which relies on Symbolic Convergence Theory. As defined by Duffy, Symbolic Convergence Theory is "a general theory of rhetoric in which groups create and share fantasies about the group and outside groups and thereby build a shared identity" (p. 293).[2] Through an analysis of selected written texts on each group's website, the author was able to determine several rhetorical narratives, including "the plea for fairness and justice," "the natural order and the resurrection of the people," and "the original people of the Earth are called to give new meaning to race" (pp. 295–305).

2.5 Discourse Analysis

McQuail (2000) describes discourse analysis as applying "to all forms of language use and textual forms," relying on the idea that "communication occurs by way of forms of 'text and talk', adapted to particular social locations, topics and kinds of participants" (p. 494). Discourse analysis of webtexts considers the socio-cultural positioning of websites; their construction of localized meaning through verbal and visual elements. Again, hate group Web texts surface as Billig's (2001) study explores the language conventions of humor as it is subverted by Ku Klux Klan "joke" websites. He found that disclaimers were used to warn against the interpretation of racist jokes on the sites as dictating violent action that the content was simply meant as a joke for those with "a sense of humor" (p. 274). However, the author concluded that much of the content was not presented as a joke, but as fact — a meta-discourse which he states "denies that the joke is a joke" (p. 278). Racist texts and images, both addressed to the fellow racist readers of the "jokes" and the ethnic targets of the jokes, provide evidence of a more diabolical discourse, one which Billig's likens to the social interactions of a lynch mob (p. 287).

Macrolevel discourse has also been analyzed on the Web, concentrating on websites which represent the voice of nation-states. In Purcell and Kodras's (2001) study of the Slovenian government's website, they approach

[2] Foss (1996) describes Symbolic Convergence Theory as being based on two primary claims: that "communication creates reality," and that "symbols not only create reality for individuals but that individuals' meanings for symbols can converge to create a shared reality for participants" (p. 122).

its texts as a response to perceived inaccuracies in the representation of Slovenia as a Balkan state; a region plagued by civil war and therefore framed as undesirable by Western countries for tourism and other commercial investments. Purcell and Kodras observe that Slovenia attempts to reinvent its identity through the texts and images on its government website, a rhetorical strategy. But they conclude that the site is ultimately part of an "an ongoing effort to negotiate [Slovenia's] position" within the "global discourse" (p. 364).

Chandler (1998) has demonstrated that, in addition to institutional websites, personal homepages may serve as a form of discourse. In his analysis of Web texts created by teenagers in Wales, and through interviews with the site creators, Chandler explores the many ways the teens conceptualize their audience and negotiate the boundary between the public and private spaces of their lives. Many of the young Web authors stated that their sites represented an attempt to communicate with visitors who might share the same interests, while others described it as something they had created just for themselves, while still others simply felt motivated to be a part of the "vast" Internet and to share their thoughts about life with others (p. 12).

2.6 Visual Analysis

Due to the multimedia capabilities of the World Wide Web, some researchers have chosen to supplement or circumvent more traditional analyses of verbal Web texts with analyses of visual texts. Website creators have employed many creative uses of graphics and images on websites, with few garnering as much browser and media attention as JenniCam.org. This personal website features a 24 h, camera's eye view of Web author Jenni Ringley's bedroom. Jimroglou (1999) uses the JenniCam site as a case study for exploring Haraway's (1991) concept of the cyborg, "a hybrid of machine and organism" (p. 149). As with the cyborg, Jimroglou observes that the boundaries between Jenni's body and technology become blurred through her camera presence online, using as an example the image of Jenni sitting in front of her computer monitor, the site of her Web cam: "The image of Jenni at her computer becomes an icon for that fusion, with her flesh melting into her keyboard" (p. 441). Feminist film theory is used to explore the meaning of JenniCam as a gendered, visual subject. Ringley's dedication to providing a glimpse into her real life includes leaving her Web cam on even when nude or engaging in sexual activity. Jimroglou recognizes the unique problem presented by these images: Jenni succeeds in breaking traditional boundaries of public and private spaces assigned to

the presentation of a woman's body while also evoking criticism from some feminists who feel she's fallen into the trap of female body objectification rampant on the Web. In either case, "Jenni's body functions as the locus of meaning, as the site of plentitude, as the root of unified meaning of JenniCam" (p. 449).

Other researchers have employed visual analysis to evaluate news media's use of Internet for presenting and distributing images. Fair and Parks (2001) present a critical analysis of images of the 1994 Rwandan genocide crisis made available by US news media via TV and the Web. Ground coverage of the crisis, in which cameras primarily captured images of frightened refugees being assisted by white, western relief workers, as well as aerial satellite images of people fleeing Rwanda en masse, are argued to have disempowered the victims of the crisis, oversimplifying the political context of the event, and distancing the viewer from the people and issues involved. Fair and Parks add that use of websites to present and distribute the Rwandan images is part of a larger trend in the domination of ownership of visual technologies by western news media, the content being produced chiefly for the consumption of western audiences, the primary users of the Internet. Fair and Parks conclude that image choices made by news media in their coverage of the Rwandan crisis, in conjunction with the mass distribution of the images through western media technologies, worked together to reinforce existing American "dislocation" from the culture and politics of Africa (p. 42).

The two studies reviewed above pay exclusive attention to images used on websites in order to render their interpretations. But rather than being mutually exclusive, the analyses of visual and verbal texts are typically integrated by researchers, as with Hesketh and Selwyn's (1999) study of the construction of school identities through images and texts on websites or Warnick's (1998) study of parodic campaign websites and their use of both verbal and nonverbal texts to advance particular political agendas. The visual design of websites, including the use and placements of images and texts on the page, provides another example of this integration of methods (Rivett 2000). A case study of the Volkswagen "Newbeetle" website discussed layout choices, such as the predominant splash image featured on the site's homepage, which the author compares to the conventions of offline magazine design. The site was also said to communicate the company's online "visual identity" through the use of a white background (p. 50). The visual analysis of the "Newbeetle" website complements a textual analysis, revealing a narrative of "alien invasion" woven throughout its pages. The VW site was then compared to the visual layout of Daniel Chandler's academic site. The latter was described as being dominated by text, emphasizing its purpose as a source of information; a ring-binder

graphic was used as the background image for the site, reinforcing its academic nature.

2.7 Ethnography

Ethnography is most easily understood as "a description and interpretation of a cultural or social group or system" (Creswell 1998; p. 58). Defining the case in ethnographic research has become a more complex process as it is performed within the study of Internet and Web cultures. The praxis of pre-Internet ethnography has been the study of single-sited, geographically centralized cultures. However, online unity and offline physical dispersion of participants in Internet cultures has required a more loosely bounded, multisited approach to ethnography (Howard 2002). The notion of the multisited community may also need to be applied to the "locations" of a community's participants online as well; providing a "thick description" (Geertz 1973) of an online community requires a more varied approach than simply observing group discourse within a chatroom or analyzing the contents of community websites. Instead, some Web ethnographers have argued that a more holistic approach is required, one in which all sites of participation and experience of community members are explored by the researcher, both on and offline (Howard 2002; Miller and Slater 2000).

Miller and Slater (2000) provide a useful example for performing a multisited model of Internet ethnography in their study of Trinidadians and Internet. The researchers visited chat rooms and websites, as well as people's homes and cybercafes, in order to capture the way Internet has intersected with the social, political, economic, and religious lives of "Trinis." Each site is described as providing a unique perspective on the culture; where chatrooms were thought to give insight into *being* "Trini" online, websites were conceived as *representations* of Trinidad online (p. 103). Miller and Slater use a qualitative content analysis to discuss the websites, pointing out the weaving together of national symbols, such as flag and map images, with the personal identities of the Web authors throughout the sites analyzed. They also discuss the reflection of political agendas in the websites, with Trinidad's most profitable tourist attraction, the Carnival festival, being a frequent focus of pride and discussion on the sites.

Researchers interested in online fan communities have found ethnography a useful approach to their studies. Bloustien (2002) included analyses of fan websites and related discussion forums in her depiction of fandom and the television show *Buffy the Vampire Slayer*. From her analyses of Web texts, as well as data collected from in-person interviews and tours of

the fans' rooms, she was able to draw conclusions about how *Buffy* fans, mostly teenagers, draw on the fantasy and magic themes of the show to maintain the imagination and "play" of youth, while entering into the more serious and "terrifying possibility of moral ambiguity" found in adulthood (p. 440).

2.8 Network Analysis

According to Garton et al. (1997), "a social network is a set of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, co-working or information exchange" (para. 2). When these social relationships are established, maintained, and/or constructed through the use or creation of websites, hyperlinks provide a means for studying the patterns of association between network members. Park (2003) identifies this type of analysis in Web research as hyperlink analysis, in which websites are conceptualized as representing individuals, groups, organizations, and nation-states, and hyperlinks between the sites as representing "relational connection[s]" (pp. 50–51). For instance, Halavais (2000) studied the hyperlinks of websites in order to better understand whether the Web is truly fostering a "World Wide" network of sites, or whether national borders are being adhered to online. He looked at 4,000 websites, examining their links to pages external to themselves, followed by an identification of the national host of the websites linked to. Through this process, Halavais was able to conclude that in fact most websites in his sample linked primarily to sites within their respective national cultures. Yet in comparison with other types of information technologies, such as ground mail and television, Halavais believed the Internet to be the most "internationalized," with more references to and from the US over international borders occurring than with other media (p. 23).

Organization-level network analysis was used in tandem with content analysis in the study of antiglobalization websites, mentioned previously, in order to understand to what degree these sites were integrated (Van Aelst and Walgrave 2002). Links between the 17 sites were analyzed using network visualization software, Pajek (http://vlado.fmf.uni-lj.si/-pub/networks/pajek/), which created a graphic map of the links between the websites. Highly connected and referenced sites were easily identified by the clustering of links to and from their location, while less integrated sites appeared as isolated outlyers with only one or two links connecting them to the rest of the network. Van Aelst and Walgrave found that the antiglobalization sites were fairly integrated, but warned that it is difficult to

assess the nature of the relationships between organizations based purely on the existence of links between sites.

In other cases, hyperlinks have been analyzed in order to uncover the social and informational networks of individuals online. Most Web browsers, such as Netscape Navigator or Internet Explorer, include a "History" function which saves the URLs accessed by a person while visiting sites on the Web. Tauscher and Greenberg (1997) used this function to analyze the browsing behaviors of 23 individuals over a six week period, looking for patterns in webpage visitation. They found that nearly two-thirds of subjects' webpage visits were ones which had been previously viewed (p. 112). In addition, the number of pages revisited frequently by participants was relatively small, further representing rather limited personal site networks among participants. Tauscher and Greenberg speculate that the "Back" function on Web browsers, which constituted 30% of their subjects' navigation actions, may be a contributing factor to Internet users' penchant for page revisitation on the Web (p. 131).

In addition to social networks the study of document networks has grown, largely due to the sheer size of the Web, which has meant that various forms of social scientific network analysis are possible. The analysis of document networks can provide opportunities for meta-analysis of content. Henzinger and Lawrence (2004) discuss methods of sampling Web pages "to automatically analyze a large sample of interests and activity in the world…by analyzing the link structure of the Web and how links accumulate over time" (p. 5186). Aizen et al. (2004) posit that "usage data at a high-traffic website can expose information about external events and surges in popularity that may not be accessible solely from analyses of content and link structure" (p. 5254).

2.9 Ethical Considerations

The importance of the need for Internet scholars to pursue research following ethical practices is clear, but the policies and means of ethical research practice are less clear. For scholars studying the Web, and particularly for those who mine archival Web materials, ethical considerations will inevitably arise during the course of research. Some material found on the Web and that is being archived is confidential, inadvertently made available and then stored by search engines like Google. The ethical positions concerning use of private material publicly archived are numerous and beyond the scope of this essay. However, it is important to note that the thoroughness with which search engines scour the Web can lead to the archiving of material that users

had neither intended to make public nor to have archived. One, perhaps overly optimistic, estimate of the speed with which search engines index the Web noted that indexing and archiving by the likes of Google outpace the creation of new Web pages (Whelan 2004). It is likely that issues that have been long discussed by Internet Studies scholars (Ess 2002) and those studying Computer-Mediated Communication concerning public versus private data will be soon at the forefront among Web studies scholars and those involved in archiving Web and other electronic materials.

2.10 Conclusion

Clearly, Web studies have come into their own. No longer in the shadow of textual analyses of text-based communities, Web research is proving that concepts such as community, culture, behavior, and meaning construction can also be effectively examined on the World Wide Web. While the wealth and variety of Web research being conducted is promising, there are still many avenues of inquiry yet to be explored.

The legacy of textual analysis of computer-mediated communication is apparent in the heavy bias toward language related analyses in Web studies. Missing are studies which explore the Web as a multimediascape, describing how images and sound are being used to circumvent verbal communication, to overcome language barriers in a global medium. With the exception of the mention of sound use on Trinidadian websites in Miller and Slater's (2000) ethnographic study, aural Web studies were completely absent from the existing body of Web research. Future projects might examine rhetorical uses of sound on the Web via analyses of commercial or political websites. Other multimedia-oriented studies could look at the ways sound and images are being used to construct identities of persons online, from individuals to organizations to national governments.

The challenges for preservation of Web content are discussed elsewhere in this volume. For Web studies it is imperative the challenges are overcome. To serve the growing number of researchers it is necessary to have Web content of all kinds, from all periods, easily accessible. Furthermore, it is necessary to be able not only to access content, but also to recreate the links within content and between sites. And, it is necessary to have available the browsers and other applications with which Web content is viewed to best understand the experience of the user. The challenges thus go beyond preservation of content and include preservation of structure of, and encounter with, content.

References

Aizen, J., Huttenlocher, D., Kleinberg, J., & Novak, A. (2004). Traffic-based feedback on the Web. Proceedings of the National Academy of Sciences of the United States of America, *101*(1), 5254–5260

Billig, M. (2001). Humour and hatred: The racist jokes of the Ku Klux Klan. Discourse & Society, *12*(3), 267–289

Bloustein, G. (2002). Fans with a lot at stake: Serious play and mimetic excess in Buffy the Vampire Slayer. European Journal of Cultural Studies, *5*(4), 427–449

Carswell, A. D. & Venkatesh, V. (2002). Learner outcomes in an asynchronous distance education environment. International Journal of Human–Computer Studies, *56*, 475–494

Chandler, D. & Roberts-Young, D. (1998). The construction of Identity in the Personal Homepages of Adolescents. Retrieved April 15, 2002 from http://www.aber.ac.uk/media/Documents/short/strasbourg.html

Creswell, J. W. (1998). Qualitative inquiry and research design. Thousand Oaks, CA: Sage

Duffy, M. E. (2003, July). Web of hate: A fantasy theme analysis of the rhetorical vision of hate groups online. Journal of Communication Inquiry, *27*(3), 291–312

Ess, C. (2002). Ethical decision making and Internet research: Recommendations from the AoIR working committee. Association of Internet Researchers. Retrieved July 12, 2004, from http://www.aoir.org/reports/ethics.pdf

Fair, J. E. & Parks, L. (2001). Africa on Camera: Television news coverage and aerial imaging of Rwandan refugees. Africa Today, *48*(2), 34–57

Foss, S. K. (1996). Rhetorical criticism: Exploration and practice. Prospect Heights, IL: Waveland

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. Journal of Computer-Mediated Communication, *3*(1). Retrieved April 4, 2004 from http://www. ascusc.org/jcmc/vol3/issue1/garton.htm

Geertz, C. (1973). The interpretation of cultures. New York: Basic Books

Graphic, Visualization and Usability (GVU) Center (1998). 10th WWW User Survey. Atlanta, GA: Georgia Institute of Technology. Retrieved March 25, 2004 from http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/.

Halavais, A. (2000). National Borders on the World Wide Web. New Media & Society, *1*(3), 7–28

Haraway, D. (1991). Simians, cyborgs, and women: The reinvention of nature. New York: Routledge

Henzinger, M. & Lawrence, S. (2004). Extracting knowledge from the World Wide Web. Proceedings of the National Academy of Sciences of the United States of America, *101*(1), 5186–5191

Hesketh, A. J. & Selwyn, N. (1999). Surfing to school: The electronic reconstruction of institutional identity. Oxford Review of Education, *25*(4), 501–520

Howard, P. N. (2002). Network Ethnography and the Hypermedia Organization: New Media, New Organizations, New Methods. New Media and Society, *4(*4), 550–574

Jimroglou, K. M. (1999). A Camera with a view: JenniCam, visual representation, and cyborgsubjectivity. Information, Communication and Society. *2*(4), 439–453

Kroeber, A. (2001). Postmodernism, Resistance, and Cyberspace: Making Rhetorical Spaces for Feminist Mothers on the Web. Women's Studies in Communication, *24*(2), 218–240

Lee, W., Hong, J., & Lee, S. (2003). Communicating with American consumers in the post 9/11 climate: An empirical investigation of consumer ethnocentrism in the United States. International Journal of Advertising, *22*, 487–510

McQuail, D. (2000). McQuail's mass communication theory (4th ed.). London, UK: Sage

Miller, D. & Slater, D. (2000). The Internet: An ethnographic approach. Oxford, UK: Berg

Milne, G. R. & Culnan, M. J. (2002). Using the content of online privacy notices to inform public policy: A longitudinal analysis of 1998–2001 US Web surveys. The Information Society, *18*, 345–359

O'Neill, E. T., Lavoie, B. F., & Bennett, R. (2003). Trends in the evolution of the public Web, 1998–2002. D-Lib Magazine, *9*(4). Retrieved March 29, 2004 from http://wcp.oclc.org/.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the Web. Convergence, *25*(1), 49–61

Pitts, M. J. & Harms, R. (2003). Radio websites as promotional tools. Journal of Radio Studies, *10*(2), 270–282

Purcell, D. & Kodras, J. E. (2001). Information technologies and representational spaces at the outposts of the global political economy. Information, Communication and Society, *4*(3), 341–369

Rivett, M. (2000). Approaches to analyzing the Web text: A consideration of the Web site as an emergent cultural form. Convergence, *6*(3), 34–56

Sarraf, S. (1999). A survey of museums on the Web: Who uses museum Websites? Curator, *42*(3), 231–243

Silver, R. C., Holman, E. A., McIntosh, D. N., Poulin, M., & Gil-Rivas, V. (2002). Nationwide longitudinal study of psychological responses to September 11. Journal of the American Medical Association, *288*(10), 1235–1244

Spooner, T. (2002). Internet use by region in the United States. Pew Internet & American Life Project. Washington, DC. Retrieved March 25, 2004 from http://www.pewinternet.org/reports/pdfs/PIP_Regional_Report_Aug_2003.pdf

Tauscher, L. & Greenberg, S. (1997). How people revisit Web pages: Empirical findings and implications for the design of history systems. International Journal of Human–Computer Studies, *47*, 97–137

Van Aelst, P. & Walgrave, S. (2002, December). New media, New movements? The role of the Internet in shaping the 'anti-globalization' movement. Information Communication and Society, *5*(4), 465–493

Warnick, B. (1998). Appearance or reality? Political parody on the Web in Campaign '96. Critical Studies in Mass Communication, *15*, 306–324

Whelan, D. (2004, 16). Google Me Not. Forbes, *174*(3), 102–104

3 Selection for Web Archives

Julien Masanès

European Web Archive
julien@iwaw.net

3.1 Introduction

The selection phase (see Fig. 3.1) is a key phase in Web archiving. It takes place at the beginning of the entire cycle and has to be re-iterated on a regular basis. Preceding the capture phase for which it provides input and guidance, it comes just after the archiving and access phase of previous crawls if any, ideally taking into account issues and necessary changes the quality review phase has raised. It comprises three phases: preparation, discovery, and filtering that will be described in this chapter.

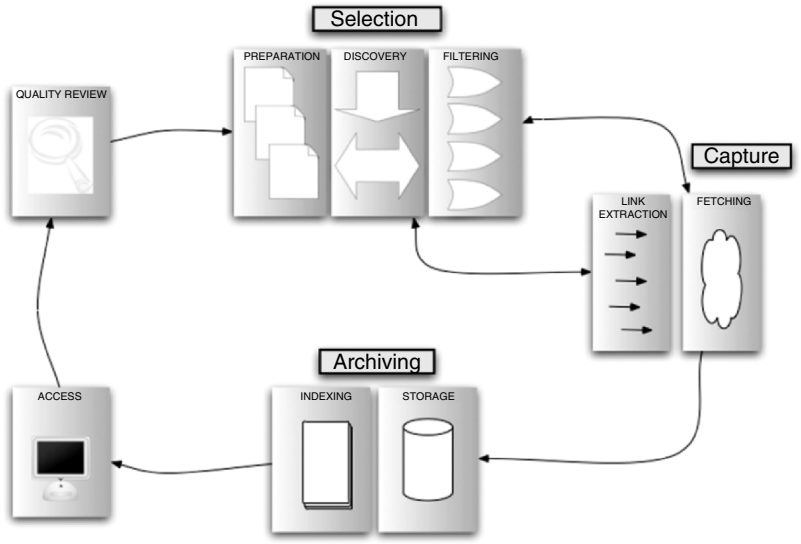


Fig. 3.1. The selection cycle, with its three phases (preparation, discovery, filtering), takes place before the capture, the archiving and quality review

The selection policy is the mark of each archiving institution. Choices made in this domain determine the type, extend and quality of the resulting institution's collection. But simply applying methods and practices developed for selection of printed material is not adequate. Web publishing is different enough from traditional publishing to require a wide revision of existing practices in this domain.

In this chapter, both the methodology and reflection on what selection means in the context of the Web will be presented with the ambition of contributing to such a revision. The chapter will cover the selection policy, the issues, and the implementation process of selection in the context of the Web.

3.2 Defining a Selection Policy

Building collections of Web material requires, when it becomes a regular activity, a general guiding document that defines the collection development policy. The benefits of defining such a policy for archiving institutions are the same as for printed material (Biblarz et al. 2001):

- It reduces personal bias by setting individual selection decisions in the context of the aims of collection building practice;
- It permits planning and identifies gaps in collection development and ensures continuity and consistency in selection and revision;
- It helps in determining priorities and clarifying the purpose and scope of each individual collection, and allows selection decisions to be evaluated by, for example, identifying what proportion of in-scope published material has been acquired;
- It can serve as a basis for wider cooperation and resource sharing.

Even if these benefits have been originally identified for collection of printed material (an electronic published resources by extension) the main principles remain for the web: avoiding personal bias or changes pertaining to a specific conjuncture hence providing continuity, defining priorities and allowing planning, positioning the collections in a larger archiving context to facilitate cooperation.

3.2.1 Target and Coverage

A collection development policy should describe at a high level and the goal driving the collection development. This comprises a description of

the context, the targeted audience, the type of access, and the expected use of the collection.

The collection's target, that is, the content to be archived, should be described in this context in general terms. This can be refined by defining inclusion's and exclusion's criteria. These criteria can be on quality, subject, genre, publishers like in traditional selection, but also on domains as defined by the Internet naming space itself. An importance difference to keep in mind for criteria adoption is their applicability on the web. It makes a huge difference in costs for instance if discovery as well as appraisal have to be made by human instead of automatically.

The concept of coverage or depth of collection has been widely used for books or serial to appraise collections, set ambitions and guide their developments. The five levels defined by the International Federation of Library Association (IFLA) are the following (Biblarz et al. 2001):

0 = out of scope
1 = minimal information level
2 = basic information level
3 = study or instructional support level
4 = research level
5 = comprehensive level

Collection can be appraised along several axis, the main one being subjects. Conspectus, by defining 24 divisions, 500 categories, and 4,000 subject descriptors can provide an outline of a collection that can be used as for systematic assessment of a library collection (Loken 1994).

The main problem when trying to apply this tool for web collections is the lack of reference to which comparing collection's completeness. This is partly due to the little number of existing institutions doing web content selection, compare to the numerous ones doing it for books or serials. For books or serials, one can easily find a large numbers of catalogs, bibliographic lists to refer to, not to mention the national bibliography made by national libraries, which provides an almost complete survey of the printed material for most countries.

This is also due to the qualities of the web as a publishing medium, which makes this type of rigid framework usually hard to apply. Traditional publishing professionalism and structuring for a well-established market prepared the ground for librarians' effort to organize printed material. As it will be discussed in section "Limitations" (implementing a selection policy) the nature of the Web makes implementation of a collection policy quite different from a traditional one.

Two main differences should be emphasized here: the connected nature of the information space and the lesser role played by professional publishers with the multiplication of content producers that goes with it.

Link connectivity deeply structures organization of information on the Web. Selection morphs from a human selection of discrete and stable units (books or serial) to a more versatile and dynamic selection of paths to be followed with certain depth and time. The topology of the Web plays a key role in both the discovery and the capture of content on the Web. It indeed tends to replace the well-structured organization of selection of the book era where publishers, collections, disciplines were natural lines along which selection was organized.

There are of course fewer differences for human-driven selection than for automated ones. However, the multiplication of content publishers, the variety of publication's forms and frequency, the nature of discovery and authority on the Web requires adapting traditional practice significantly.

For an example of a selection policy closest as possible to the traditional model, see the NLA's selection policy (National Library of Australia, 2005).

3.2.2 Limitations

Whereas traditional acquisition policy had mainly to deal with financial limitations (for acquisition, processing, or storage) web archiving is also directly and permanently hindered by technical difficulties for capturing content. Different types of technology challenge current capture techniques: the hidden web (see Chap. 5 on hidden Web archiving by Masanès 2006a), streaming content, highly interactive content etc. There are hence hard limits to what can actually be archived. There is also an inherent limitation to server-side archiving (the main archiving methods) that can only capture functionality that are supported by client-side code. The development of AJAX web programming style based on content exchange between the page and the server without reloading the page can augment the amount of material not captured by crawlers. These limits have to be included in a selection policy when possible, as they will impact the resulting archive quality. Here is for example the list of exclusion of NLA's policy. Although we could not say whether this had been the main or only one of the reasons behind some exclusions (cams, datasets, games), technology would have been a challenge in these cases anyway.

- Cams (websites employing a Web camera that uploads digital images for broadcast);
- Datasets;
- Discussion lists, chat rooms, bulletin boards, and news groups;
- Drafts and works in progress, even if they otherwise meet the selection guidelines;
- Games;
- Individual articles and papers;
- News sites;
- Online daily newspapers for which print versions exist;
- Organizational records;
- Portals and other sites that serve the sole purpose of organising Internet information;
- Promotional sites and advertising;
- Sites that are compilations of information from other sources and are not original in content;
- Theses (the responsibility of universities and the Australian Digital Theses Project).

3.2.3 Gathering Patterns

Building Web collection can either be done on a continuous basis or through campaign or snapshots.

Examples of these campaigns are elections sites acquisitions or snapshot domain crawls. Although archiving campaigns can change on emphasis or thematic, they should be done accordingly to the collection development policy. Conversely, the collection development policy should describe, when possible, the campaign patterns to make sure the end result is consistent with the overall aim of the collection development. A campaign pattern description should at minimum comprise the trigger(s) (calendar, events, etc.), duration(s) of the campaign as well as the possible bridges to be established between campaigns.

Here is a very simple example of such campaign pattern description:

Start national domain snapshots every three months, with a campaign duration of 60 days and by using as entry points, list all the domains found in the previous campaign.

3.3 Issues and Concepts

3.3.1 Manual vs. Automatic Selection

A recurrent theme in the literature on Web archiving is somehow simplistic opposition between manual selection and bulk automatic harvesting allegedly considered as unselective. The former is misleadingly supposed to be purely manual whereas the latter is similarly falsely considered as comprehensive. We prefer to insist on the fact that Web archiving always implies some form of selectivity, even when it is done at large scale and using automatic tools.

There are two levels at which this selectivity and the determinism of automatic tools takes place: discovery and capture of material. Comprehensiveness as opposed to selectivity is a myth as Web's size and versatility make it impossible to discover and to capture every possible instantiations of content for all possible readers. Actually, there is a default selectivity of large-scale crawlers in term of the extent, depth, and time at which they crawl sites, all these in turn being dependent on resources used, capacity to extract links, queuing method, politeness to servers, entry points used, etc. We prefer to use in this book the term holistic archiving, defined as archiving made by open crawls using link extraction for discovery.

On the other end, manual selection of Web documents rarely happens without requiring utilization of automatic discovery tools like search engines. See a detail modeling of user/machine interaction for IR in general and Web search in particular in Ellis et al. (1998). It should also be noted that these tools add an access bias (ranking methods) to the crawl bias (see for instance Introna and Nissenbaum (2000) and Vaughan and Thelwall (2004).

And even in the case where discovery would entirely be done manually, capture is most of the time done with tools based on link extraction (site copiers). Here again, one has to be aware of the fact that these tools always have at least embedded capture bias like definition of scope, implicit or explicit exclusions of content by format type, prioritization of capture, resources constraints (hardware, bandwidth), etc. This underlying determinism of web capture has a sufficient impact on the final resulting collection not to be underestimated (see Chap. 4 Roche 2006).

3.3.2 Location, Time and Content

As we all know, and despite the fact that this is totally counter-intuitive, there is no such a thing as reference to objects on the Web. URLs provide references to locations, not objects and applying a selection policy in a space only structured in terms of location is more challenging than one could think at first glance. To take a familiar example, it is like walking in an open stacks library and selecting books only by their location (shelves) while objects can be moved or removed from time to time. But going straight to the shelf containing medieval history books, if this is what one is passionate about, is only one of the two main possible means for a reader of selecting book in a library. Most of the time, she/he would use the other one, the catalog. And catalogs handle objects identifiers, for which they provide location.

Web references on the contrary, handle locations first, then objects. The difference does not only come from the order, but also from nature of this relation. In one case (the library), the relation between the object and its location is maintained by the library itself which guarantees it will work whatever happens to the original publisher (and we know that libraries in general last longer than publishers). Whereas on the Web the relation between the object and its content depends on the publisher. Moreover, it actually also depends on the publisher's technical ability and permanent use of resources needed to serve content online. This is what can be called permanent publishing.

It would be misunderstanding the real nature of the Web to think of this characteristic as a shortcoming. It indeed offers the ability to create a space where each defined location (like a domain name) is a source of possibility instead of being a placeholder for fixed content, where the producer will have the possibility to propose, update, change, and remove content, which is the very nature of publication after all. A location is then a way to connect to someone's or something's stream and this can become a criterion for selection.

A consequence of this is the introduction of a new temporal dimension in the archiving process. Actually, archivists are more used to deal with this than librarians are. Their activity has always been closely linked to the document lifecycle whereas librarians have mostly been working on stabilized (published) content. For the Web, as contents update or removal can occur at any time with great facility, this temporal dimension has to become a core component of any process of archiving. The fixation of a particular state of content, which used to be done only by the publisher, is now also made by the archive.

This is a new responsibility for preservationists in general, and it alters significantly the usual relation between location time and content in traditional archiving as well as actor's roles in this process. By capturing and storing content on its servers, the Web archives removes completely the resources dependency to its original creator and eliminates the virtuality of any Web location, "freeze" the resource in time and therefore reduces its dual nature to its content aspect only. This, which used to be done by publishers and printers for printed material, falls under the responsibility of the archive for Web material.

3.3.3 Hubs and Targets

Even if little has been done on selection with archiving as a goal, a significant precedent effort was made for creating reference list or subject gateways pointing to selected resources on the Web. It is interesting to note that the dual nature of Web references is sometime pulled on one side or the other. The location aspect is certainly more important for resources that contain themselves mostly reference and links to other resources (hubs, also called webographies, etc.). This is also the case where the resources are valued for providing up-to-date information. They are valued and selected as locations where updated and reliable information on a subject can be found, that is, locations where change can and has to happen.

Some resources on the contrary are valued for their content itself. This is usually the case for smaller piece of content, pages, single documents, and/or content dependent on a specific time (events).

The graph theory offers concepts that can prove useful to characterize the two types. Considering the graph formed by pages (nodes) and links (directed edges) it is easy to calculate the in-degree of a node (number of nodes that link to it) or its out-degree (number of nodes it links to). From a pure structural point of view, we can expect a hub to have a high out-degree, and a low ratio content/out-link (content can be measured by the quantity of information the node contains for instance). On the contrary a target can be expected to have a low out-degree and overall an important ratio content/out-link. If this target is important, it can be expected to have, in addition to this, a high in-degree.

An iterative definition of hubs and authority can be found in Kleinberg (1999) and can be summarized as:

Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

Both concepts can be applied for Web archiving with slight modification. We will use the term "target" instead of authority in this book, in order to remain neutral with regard to the "authority" of each individual node selected. An archiving policy can target content for a wider range of reason than just authority. Important to note is the nonexclusivity of the two types: a hub can also be a target for a selection policy as it can provide content as well as referral information on other resources. Hubs and target can be selected and linked to by subject gateways. Targets are vulnerable to time as time entails, at least potentially, change, whereas hubs are pointed to by subject gateways because of the same reason, as they offer a good chance to be changed if needed (updated).

One could conclude rapidly that hubs are of little interest for Web archiving selection policy, which should concentrate its effort on the first type, vulnerable to time. But this is not the case, at least for two reasons. The first one is that hubs can attest of relations like valuation, endorsement, etc. that could prove to be of great interest and deserve archiving for themselves. The second reason is that even when they are not targets, they certainly are a mean for finding targets. Hubs are indeed a tool of choice to implement a web archiving policy.

3.3.4 Entry Point and Scope

The term hub and target have been defined from a pure structural and content-quality perspective. From a practical point of view, we need to introduce two other terms related to how a selection policy can eventually be implemented. The first one, entry point (EP) also called seed, could be confused with the concept of hub. An EP is defined as the first node from where path to other documents will be found in a crawling process. As they both have to do with out-linking, hubs, and entry point often tend to be confused and indeed, most EP are usually hubs themselves. But this is not always the case. An example of this is the site's homepage, often considered as the EP for a site crawl whether it is a good hub beyond the site itself or not. The related concept of scope can be defined as the extent of the desired collection, delimited by a set of criteria. Criteria can be topological (the Italian domain), thematic (sites related to biology), based on genre (blogs) time (site stale since the last 2 years), etc. Each time a new path is discovered from the EP or pages linked from it, it has to be evaluated to see if it fits in the scope or not. To be operational, scopes have to be defined in a way that enables direct and possibly automatic verification. If not, a systematic human evaluation is needed for each new link discovered. If the selection policy is applied at the site level, this will only be necessary when a link to an external site is discovered.

3.3.5 Level of Application

Traditionally, selection policies have implicit application levels, defined by the physical shape of their object (obviously, a selection policy applied at the level of book pages would have been nonsense). Stating the accepted types of documents (books, serial but not reports for instance) was sufficient in this environment. With the Web, this is no longer the case. There is no obvious level of selection and sometimes, levels are difficult to delineate (see for instance Thelwall 2002; Halavais 2003). It has been argued that the appropriate level for analysis in political science is the Web sphere, defined not simply as collection of websites but as

A set of dynamically defined digital resources spanning multiple websites deemed relevant or related to a central event, concept or theme, and often connected by hyperlinks. The boundaries of a Web sphere are delimited by a shared topical orientation and a temporal framework" (Schneider and Foot 2005).

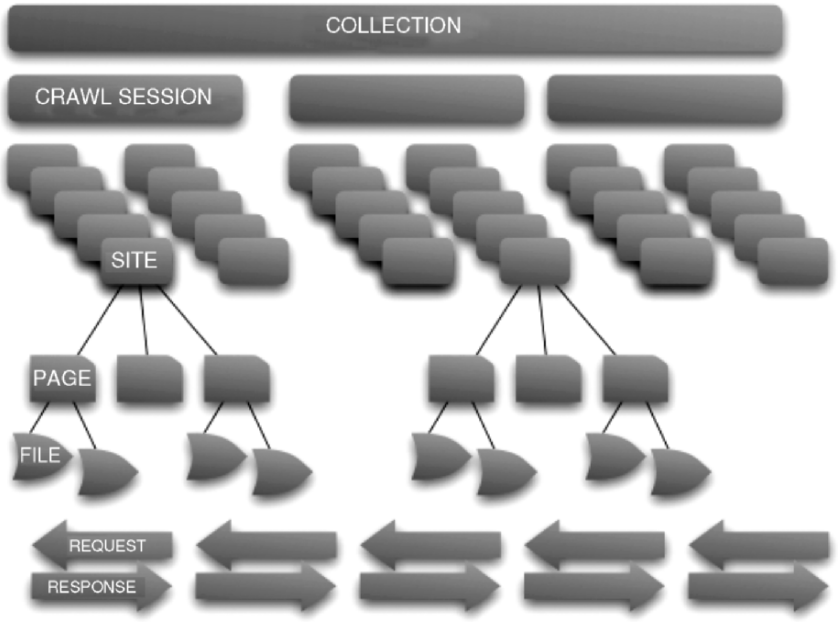


Fig. 3.2. Levels of information in Web archiving, from the request/response exchange to the site level for Web levels, from the crawl session to the collection for the archive's level

For selection, at least two working levels have to be considered: the page and site level. The page level corresponds to the immediate experience of web users and can therefore always be specified by human as well as by tools (browsers). It includes the skeleton (usually an html page) together with its embedded elements when rendered by the browser that is, images, style sheets, script files, etc. It can also consist of a non-web document and be rendered using helper applications (like a PDF document for instance).

Although it is used in many research studies (McMillan 1999), the site level is more difficult to define. The intuitive notion of a site refers to a related set of resources, sharing a same creating entity. The network notion refers to the host or web server that is, the machine serving the content of the site. Finally a purely topological notion of a site can be defined as a naming space section (a domain name for instance).

Confusion arises from the fact that these three levels are not clearly delineated internally. An entity can be an organization, a department, a person or even a project with its own site. A single machine can host several websites (and conversely, a site can be hosted on several machines). There is neither a strong naming convention for sites. A site can be located at the level domain (netgramme.org) at the sub-domain level (zone.netgramme.org) or even at the directory level (netgramme.org/blog). The situation worsens as the three levels get confused one with the other. This flexibility in possibilities (and in actual usage) is typical of the web.

As selection policies are usually content driven, they should focus on the first and third level (discarding the machine level). Defining target sites can be done by assessing in each case the appropriate entity level. If the collection targets neuroscience related material, neurobiology research laboratory's sites are certainly more interesting than entire university Websites for instance. The hierarchical nature of naming conventions can often help here as long as they have been applied for the site construction. Therefore when identified, the most characteristic path can be passed over to crawler that can deal with it by getting only the content further down in the hierarchy. For instance, if the site has been identified under a specific directory (netgramme.org/blog) than the crawler can limit the crawl to content that is under this directory (however, deep). Note that the naming hierarchy goes from right to left for the domain name and left to right for the path (see Fig. 3.2).

3.4 Selection Process

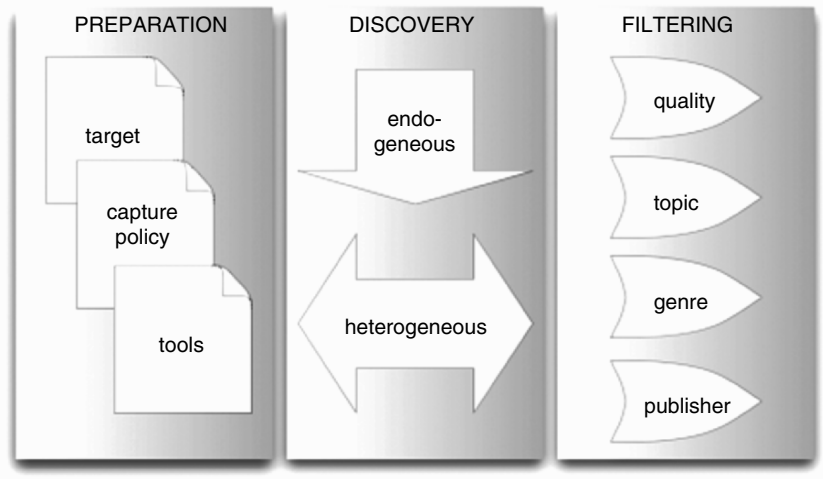


Fig. 3.3. The phases of the selection process: (1) preparation with its main output (the target definition, the capture policy and the list of tools to be used), (2) endogeneous and heterogeneous discovery and, (3) filtering according to quality, topic, genre or publisher

The selection process can be divided in three main phase: preparation, discovery, and filtering (see Fig. 3.3). Although these phases can occur in a sequential order or can be mingled together to some extent, we will present them as logically distinct for sake of explanation.

3.4.1 Preparation

This phase is a key for the success of the whole process and should not be underestimated in terms of time as well as resources required to perform it successfully. The main objective of this phase is to define the collection target, the capture policy and the tools for implementing it.

For topic-centric as well as domain-centric collections, input here is mainly required from domain experts (references, librarians, archivists, scholars) that have to define what the target information space is, how it can be characterized in extension and granularity, and which frequency of capture will be applied.

The definition has to be precise enough to be implemented. Whether the discovery and filtering phases will be implemented manually or automatically makes a huge difference with regards to what “precise” means here.

Here are two examples, one with a strong human input in the discovery phase, the other where crawling will entirely perform discovery.

Example 1: Presidential election campaign collection:

- The archiving campaign will start 3 months before and end 1 month after the election date;
- All party, campaign, blogs sites of each official candidate will be archived entirely each week during the capture campaign;
- Main analysis, commentary and humorous website entirely dedicated to the elections will be archived every month during the capture campaign;
- Individual articles from the national and regional newspaper's websites will be archived once;
- The presidency website will be archived each month during the archiving campaign.

Example 2: National domain capture:

- Capture all French public sites on a bi-monthly basis, French sites being defined as sites from the .fr TLD or sites in generic TLD that are hosted by servers in France (based on the telephone number provided for DNS registration);
- A seed list of 12 general directories is provided to initiate the first crawl. Next crawls will start from the list of sites discovered during the previous crawl with complementary list of missing sites manually selected.

It is quite obvious that appraising what a campaign site of a candidate is, as in the first example, requires human judgment. Finding as well as filtering sites along this criterion will require manual processing. Whereas, in the second example the discovery and filtering criteria can be directly interpreted by robots. The preparation phase also requires defining which tools will be used during the discovery process. Four categories of tools can be used:

3.4.1.1 Hubs

Hubs can be global or topical directories, sites or even single pages with important links relevant to a given subject. These hubs are maintained by human, and often provide a valuable source for identification. Their reliability, freshness as well as their coverage has to be assessed on a periodic basis. When possible, direct contact with the person(s) in charge of a hub can be fruitful to better understand how their input can be used. Monitoring these hubs during the capture campaign as is necessary to ensure, they

remain relevant and exploits their input. This can be facilitated if they provide RSS or Atom threads.

3.4.1.2 Search Engines

Search engines can facilitate discovery of relevant material as long as precise enough query terms can be defined. Utilization of specialized search engines, when possible, can greatly improve relevance as well as, sometime, freshness of results. When the topic is closely related to a specific event, one should expect search engines to find relevant information only with a certain delay, which limits their usefulness in this type of capture. It can be helpful to define a list of queries as well as a list of search engines to use during the preparation phase. A periodicity of query and/or a mechanism to get updates (query-based RSS feeds or agent that filter new results) is also worth defining.

3.4.1.3 Crawlers

They can be used to extract links from already known material in a systematic manner. This can be used for exploring proximal environment of a given set of EP.

3.4.1.4 External Sources

Non-Web sources can be anything from printed material to mailing lists etc. that can be monitored by the selection team. They should be used when possible as they often provide fresh resources as well as different directions for the collection. Here again a monitoring process as to be put in place as this can easily become time-consuming and yield too little compared to time invested. It should be noted that, depending on the external sources authority, an item's citation by this source could, by itself, become a reason to select it.

At the end of the preparation phase, the following output should be available:

- The collection's target description;
- The capture policy, including the level of application, the frequency and extension of the capture;
- The list of tools that will be used for discovery and capture with a description of how they will be used.

3.4.2 Discovery

The main goal of this phase is to determine the list of entry points that will be used for the capture as well as the frequency and scope of this capture. It should be noted that there is a quite clear cut between discovery and the crawl itself for collection done manually, even if the list of entry points can be updated based on links discovered during the crawl. For automatically built collections, this difference is blurred by the fact that most of the discovery occurs during the crawl itself by links extraction. However, we can differentiate for both methods, "endogenous" discovery made from the exploration of EP's and crawled page's linking environment, from "exogenous" discovery that will result from the exploitation of hubs, search engines, and non-Web sources. Where "manual" collections mainly rely on exogenous discovery, "automatic" collections rely mostly on endogenous discovery for building collection.

Endogenous discovery takes advantage of the link structure of the Web to traverse and find new material. There is evidently a good chance that sites or more generally resources linked together deserve to belong to the same collection, as links are usually the expression of a semantic or topical relation (on the centrality of links see for instance Jackson, 1997). We will see in the next section (Filtering) how to qualify this topical proximity using textual content or linking structure. Let's just note here that related content can sometime be connected not directly but through several hops. This is for instance the case in competitive environment. Competitors will hardly directly link each other. Traditional citation analysis has studied this phenomenon extensively and showed that utilization of co-citations (two papers with no direct reference to each other will be both linked from or link to a common third paper) is a way of overcoming this problem (see for instance Garfield 1979).

The same can occur also at a macrolevel, the community level, where communities are strongly interconnected (which permits good discovery within the community), but loosely if at all connected across communities. Thus, the community forms a closed sub-graph where an endogenous discovery process can be trapped. In this case, it is either necessary to permit several hops with no filtering rules applied to "tunnel" out of these sub-graph (Bergmark et al. 2002), or use heterogeneous strategies (like insertion of meta-search results (Qin et al. 2004) in the discovery process). Finally, let us note that from a discovery point of view, linked resources belonging to different websites bear more value than those within the same site. They indeed bridge different information space, possibly belonging to different publishing organizations. Eiron and McCurley show that 33% of links are of this type (Eiron and McCurley 2003).

Heterogeneous discovery does not share this problem of sub-graph trap, as it uses sources that are not (or supposed not) to be linked to any specific community or portion of the hypertext graph. It, however, entirely depends on the type, quality and usability of the sources used: hubs, search engine or non-Web sources. The usefulness of the first type (hubs) obviously depends mainly on the quality and freshness of the source. Using the second one (search engine) permits to exploit the large and neutral Web exploration artifacts that giant search engine crawls represent. The difficulty is then to be able to query efficiently their huge inverse index. However, it has been shown (Lawrence and Giles 1998) that search engine coverage is relatively small and that there is little overlap between them. Using several of them (directly or through meta-search engines) is required to achieve a better coverage.

Non-web sources require specific monitoring, adapted to each case. The paper press, for instance, can be a rich source for an event-oriented collection either directly when websites are mentioned in articles, or indirectly when names or specific words can be found and use for a search.

Heterogeneous and exogenous discovery are not completely separated and a blend of the two can result in better results (Qin et al. 2004).

When entry points are discovered, a frequency and a scope of capture have to be assigned to them. This can be done individually or based on grouping of EPs. It is usually either done at the collection or capture campaign level, by defining one or several profiles of captures.

The usual frequencies are "once only", weekly, monthly or every *x* months. It is rare that capture have to be done on a daily basis or even several times a day. This can however be necessary for online news sites (see for instance Christensen-Dalsgaard 2004) or sometime for event-related captures like the September 11 Web Archive (Schneider et al. 2003).

The scope of capture is also important to define. As mentioned earlier ("Level of application") defining boundaries in the web is not simple. However, the page and the site level (defined as the domain, sub-domain or directory location) can be used, as they are directly understandable by crawlers. The units can be either in the entry point list or discovered from them. They can be used for defining the boundaries of capture in a restricted or extended manner. The restricted scope is to limit the capture to a specific page or sites that

Following any links from the entry points could result in an endless crawl of the entire Web. It is therefore necessary to shape this discovery process accordingly to the selection policy.

3.4.3 Filtering

The filtering phase's main goal is to reduce the space opened by the discovery phase to the limits defined by the selection policy. As already mentioned, if this phase can be logically distinguished from the others and particularly the previous phase of discovery, they can, in practice be combined one with the other.

Filtering can be done either manually or automatically. Manual filtering is necessary when criteria used for the selection cannot be directly interpreted by automatic tools or robots. This can be the case when high level characterization, subjective evaluation, and/or external knowledge is needed. As costs associated with individual selection and updating of list of resources by humans are high, there is a strong incentive to find ways of replacing or enhancing the efficiency of this scarce resource. Furthermore, the frontier between what is and is not interpretable by robots is highly dependent on technological evolution, and significant progress can be expected in this domain. It is, after all, a question of exploiting humanly generated intelligent information like words and links in pages in an automatic manner. Strong correlation between human appraisal and what can be assessed from structural properties reflecting this collective intelligence of the web have been established for instance by Masanès (2002) in the context of web archiving.

But in some cases, human input is still needed. Consider for instance our first example of collection policy, defining what the "main analysis, commentary and humorous website" requires both high-level characterization ("analysis", "commentary", "humorous") and subjective evaluation ("main"). "All party, campaign, blogs sites of each official candidate" also requires knowing which are the official candidates, that is, an external knowledge about the campaign itself. When direct human input cannot be replaced, it can be greatly optimized by using an appropriate and ergonomic presentation of items to be filtered. This includes for instance contextual information, visualization tools, and maps (see Cobb et al. 2005, for instance).

It is also important to define the appropriate level at which manual filtering has to occur to avoid duplicate evaluation (the higher the better to save time).

Several evaluation axes can be used, alone or in combination, for manual selection:

3.4.3.1 Quality

This comprises an appraisal of authority and credibility for secondary resources (like, in our example analysis and commentary websites of the campaign) and relevance and authenticity for primary resource (sites of political party).

3.4.3.2 Subject

A subject can be delimited along the traditional scholarship disciplines (biology, geology, etc.), or according to a specific event, person or organization, or any object in general, which will be envisaged from various points of view (like in our example, the elections). Here again, primary as well as secondary websites are to be considered for inclusion in the selection policy.

3.4.3.3 Genre

Web genre is institutional website, blogs, personal pages, forums etc. This can either be the main selection criteria for genre studies or an additional criterion (as blogs in our example). Genre have been studied in the context of the Web to see how they replicate or diverge from the genre in the printing world (see Crowston and Williams 1997), or how they can be automatically identified (Rehm 2002) and several genres have been studied like newspaper (Eriksen and Ihlström 2000), in homepages for instance Ryan et al. (2003) or FAQ (Crowston and Williams 1999).

3.4.3.4 Publisher

Traditionally publishers' reputation or specializations have been used to guide selection of printed material. It is often difficult to determine the publisher of a website and only very regulated top-level domains TLDs offer homogeneity of publisher's type, like the .mil for US military sites. Using the publisher or site owner as a basis for selection hence requires most of time detailed analysis of the site as no guarantee exists that claims of identity are legitimate on the Internet. The DNS information can be used to find out who is renting the domain as it requires registration of name and contact information of the technical and administrative contact. But depending on the TLD and the way it is managed, this can be either well quite complete or very limited.

3.5 Documentation

Whatever criteria are used for manual or automatic selection, it is necessary to document carefully the selection process. As we have seen previously (see Chap. 1 of this book Masanès, 2006b), web archiving can only achieve sampling of instantiation of content. As time goes, the original context of the sample is lost and no clue will remain for researchers to understand what the archive represents. To limit this, it is absolutely necessary to document each aspects of the selection process in order to provide elements of assessment for the future. This has to be done for the various phases outlined in this chapter (preparation, discovery, and filtering).

For the preparation phase, the main aspects to document are:

- The target;
- The capture policy and infrastructure (this comprises the technical capacity, software used, priority, politeness etc.);
- The tools used (name, regularity and context of use, staff, etc.).

Documenting the discovery and filtering is even more important as this will "tell" why a piece of content is or is not is in the collection. When possible, this should be documented at the item level. Keeping a list of URI that were discovered but filtered out can for instance be useful to later understand how the collection was built and therefore, what it represents compare the live web. How endogenous discovery was used is also important to document to be able to reconstruct path that were followed and map those that were not.

3.6 Conclusion

Selection is a key issue for web archiving. Manual selection can prove useful for a specific community and/or goal, where high-level assessment of items is necessary. As long as they cannot be made by robots, human selection has to be used, and it is necessary to organize it optimally. But even for holistic crawls, there is a level of selectivity and prioritization that has to be acknowledged and organized. We have presented in this chapter an analytical view of this process selection that tries to show that the two approaches share the main elements of this process, even if their relative importance and their use are different.

References

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002). *Focused crawls, tunneling, and digital libraries*. Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries, Roma, Italy

Biblarz, D., Tarin, M.-J., Vickery, J., & Bakker, T. (2001). Guidelines for a collection development policy using the conspectus model. *International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development*

Christensen-Dalsgaard, B. (2004). Web Archive Activities in Denmark. *RlG DigiNews*, *8*(3)

Cobb, J., Pearce-Moses, R., & Surface, T. (2005). *ECHO DEPository Project*. Paper presented at the 2nd IS&T Archiving Conference, Washington, USA

Crowston, K. & Williams, M. (1999). *The effects on linking on genres of Web documents*. Paper presented at the 32nd Hawaii International Conference on System Sciences (HICSS-32), Hawaii, USA

Crowston, K. & Williams, M. (1997). *Reproduced and emergent genres of communication on the World Wide Web*. Paper presented at the 30th Annual Hawaii International Conference on System Sciences (HICSS-30), Wailea, USA

Eiron, N. & McCurley, K. S. (2003). *Locality, hierarchy, and bidirectionality on the Web*. Paper presented at the Workshop on Web Algorithms and Models

Ellis, D., Ford, N. J., & Furner, J. (1998). In search of the unknown user: Indexing, hypertext and the World Wide Web. *Journal of Documentation*, *54*, 28–47

Eriksen, L. B. & Ihlström, C. (2000). *Evolution of the Web news genre – The slow move beyond the print metaphor*. Paper presented at the 33rd Hawaii International Conference on System Sciences (HICSS-33), Hawaii, USA

Garfield, E. (1979). Mapping the structure of science. *Citation indexing: Its theory and application in science, technology, and humanities.* NY: Wiley.

Halavais, A. (2003). *Networks and flows of content on the World Wide Web networks and flows of content on the World Wide Web*. Paper presented at the International Communication Association, San Diego, USA

Introna, L. D. & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *Information Society*

Jackson, M. H. (1997). Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication*, *3*(1), http://www.ascusc.org/jcmc/

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computer Machinery*, *46*, 604–632

Lawrence, S. & Giles, C. L. (1998). Searching the World Wide Web. *Science*, *280*, 98–100

Loken, S. (1994). The WLN Conspectus. *Cooperative collection management: The conspectus approach* (p. 107). New York: Neal-Schuman

Masanès, J. (2002). Towards continuous Web archiving: First results and an agenda for the future. *D-Lib Magazine*, *8*(12)

Masanès, J. (2006a). Collecting the hidden web. In J. Masanès (Ed.), *Web archiving*. Berlin Heidelberg New York: Springer

Masanès, J. (2006b). Web archiving: issues and methods. In J. Masanès (Ed.), *Web archiving*. , Berlin Heidelberg New York: Springer.

McMillan, S. J. (1999). *The microscope and the moving target: The challenge of applying a stable research technique to a dynamic communication environment*. Paper presented at the 49th Annual Conference of the International Communication Association (ICA-99), San Francisco, USA

National Library of Australia. (2005). Online Australian publications: Selection guidelines for archiving and preservation by the national library of Australia

Qin, J., Zhou, Y., & Chau, M. (2004). *Building domain-specific web collections for scientific digital libraries: A meta-search enhanced focused crawling method*. Tuscon, AZ, USA

Rehm, G. (2002). *Towards automatic Web genre identification a corpus-based approach in the domain of academia by example of the academic's personal homepage*. Paper presented at the HICSS-35

Roche, X. (2006). Copying websites. In J. Masanès (Ed.), *Web archiving*. Berlin Heidelberg New York: Springer

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

Schneider, S. M., Foot, K., Kimpton, M., & Jones, G. (2003). *Building thematic Web collections: Challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive*. Paper presented at the 3rd Workshop on Web Archives (IWAW'03), Trondheim, Norway

Schneider, S. M. & Foot, K. A. (2005). Web sphere analysis: An approach to studying Online action. In C. Hine (Ed.), *Virtual methods: Issues in social science research on the Internet*. Oxford, UK: Berg.

Thelwall, M. (2002). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of the American Society for Information Science and Technology*, *53*(12), 995–1005

Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing and Management*, *40*(4), 693–707

4 Copying Websites

Xavier Roche

xavier@httrack.com

4.1 Introduction – The Art of Copying Websites

The fundamental difference between copying file structures or ftp sites, and websites, lies on the very deep nature of the World Wide Web. No "directory listing" in the HTTP protocol, nor bulk transfer of website zones: it is a design choice for the Web – a collection of heterogeneous resources, not necessarily related to each other. A collection of pages generated from a database content, for example, is an unstable realm of information – shall the database be flushed, it then disappears. The Web is fundamentally a moving form: ftp directories often change, but you can easily synchronize them to obtain a up-to-date state wherever you are. It is only a matter of data stored on a file repository. But a Web page is potentially unique: it can be a clock counter, a real-time information delivered on demand, a user-specific or session-specific view of a more general data collection. It can be anything you want: its internal logics are hidden to the navigator and its user. Simply speaking, an ftp server is a collection of files, more like a remotely accessible, public hard disk. A Web server is more a collection of logical resources delivering content to clients. These logical resources can be programs connected to databases, to other systems, with complex interactions with the user's preferences and needs and the server environment (database, external sources, current state, etc.). The remote client never sees this logic: only the resulting content is accessible.

Hence, there are three ways to copy entire websites (see Chap. 1 'Web Archiving Issues and Methods'): the first, server-side archiving is the hard one. It consist in contacting each Web masters and convince them to organize a copy of their internal information system files, database schemas and system specifications, and then set-up the same architecture – hardware, software, environment (such as external data sources that could be used). This solution, generally hard to deploy even for the Web masters themselves, cannot be seriously considered for a wide-range solution. The

second choice is to have this done close to the server and record all transactions (transactions archiving). The last one is to automatically collect the delivered information directly from websites, as a regular browser would do (client-side archiving).

It is a makeshift as mirrors will never be perfect: as taking a photograph of a moving scene, you will not be able to recreate its movement. You will not be anymore able to get the real-time feeling when browsing online temperatures reports or stock exchange movements. But it is an acceptable compromise in term of feasibility and quality in most cases – a static photography of a website, a photography that we would preserve in a photoalbum. A photography that could be viewed again and again without even bothering about the existence of the live model anymore.

Copying websites using this technique is something very intuitive: the method is the exactly the same as if you were copying a website by yourself, using a regular browser. You would start from the first page, save it, save the associated images, and then click on each links to view them, save the corresponding pages on disk, and carry on until you saved all the pages you wanted to copy. After that, make some changes inside the HTML pages so that they can be viewable locally by your browser, checking all relevant tags. But copying more that one or two pages manually is a bit tiresome, and an automated tool can be a relieving solution. The automated link enumerator is generally called the "parser," and the automated remote data downloader the "crawler." These two main components have additional roles: the parser is also responsible for ensuring that links will still work in a local copy, by changing the URL syntax to a compatible, a "fully relative" one; and the crawler is also responsible for handling caching and updates.

There are many reasons why you want to copy a website. At the national school of engineering of Caen, we wanted to archive small and medium websites, not for classical archiving, but to gather technical sites run by individuals which were moving very quickly. We also wanted to collect sites with large multimedia contents that were unreachable using the existing domestic dialup lines, store them on permanent medias (such as CD-Rom), and view them offline. In general, we needed a tool to collect very specific information for end-users from the WWW.

The HTTrack project was born to fulfill these needs: an easy-to-use tool that would allow regular users to make copies of small – but important – parts of the World Wide Web. Its design was rather experimental: Internet and related network architecture were fairly new domains we were discovering – and in particular, website copying was a totally new subject for us. The experience acquired through the development of this project will illustrate the method – "the Art" – of copying websites, and the suggested solutions for the multiple drawbacks encountered.

4.2 The Parser

4.2.1 The HTML Core Parser

The HTML parser is one of the two core components in a Web copying tool. Given an HTML[1] page data – that is, essentially an 8-bit[2] text file composed of plain text and markup tags – and its associated information such as the original URL,[3] the HTML parser's goal is to scan the page to collect links, analyze them and pass them back to the crawler. The HTML structure is not relevant to collect links: we are primarily interested in a limited number of tags, such as "a" or "img" elements, that will potentially contain hyperlinks to other resources (images, style sheets, HTML pages, etc.). Their position inside the page is generally not important, except in specialized domains where advanced heuristics can attach additional information such as the theme being discussed "around" these tags, allowing to follow certain pages and not other (irrelevant) ones. For a regular parser, the only useful information is the tags and their embedded properties.

The simplified core automaton is fairly easy to understand: a linear scan of the HTML page data bytes, starting from the beginning, detecting starting tags (<) and recognizing the various HTML elements by their names (see Fig. 4.1).

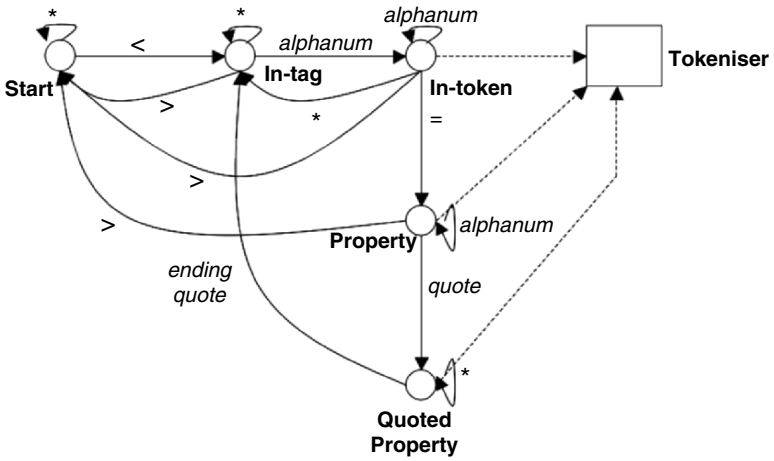


Fig. 4.1. Core parsing automaton

[1] See [1866].

[2] Note that the page character encoding will be important for link naming, especially on UCS2 file systems (including Windows ones).

[3] See [1738].

There are two classes of items to recognize inside HTML tags: tag names, such as “img” or “a”, and tag properties, such as “href ” or “src”. We can split these tags in two main groups: tags that allow to embed resources (such as images or style sheets loaded in the current page), and tags that allow to navigate to other resources (hyperlinks). For a given page, you can skip irrelevant links from the second group (e.g., links beyond the scope of the mirror) – the links will be unreachable in a disconnected (offline) environment, but this will not change the page aspect. But you have to be more careful concerning the first tag group, or the page will not be properly viewable when disconnected. In particular, you may have missing images or a totally broken page layout due to missing elements, such as style sheets or embedded scripting files. Hence, link URLs are not the only information that will have to be passed back to the crawler: the “tag context”, such as whether it is an “embedded” resource or not, will also be important to take the decision “take this link or not.”

Links themselves will be extracted by the tokenizer by analyzing well-known properties, which will be converted in their absolute form[4] using the original page URL: a protocol part, “`http:`,” a host authority “`//www.example.com`,” and a relative path “`/index.html`.” For example, the relative[5] link “`/top.html`” inside the page “`http://www.example.com/foo/index.html`” will be converted into the link “`http://www.example.com/foo/top.html`.” The link position will then be checked to ensure that it fits the default scope of the mirror; the check consists of a regular expression which value is by default the main URL prefix. If we started the mirror from “`http://www.example.com/foo/index.html`,” the default scope would be “`http://www.example.com/foo/*`” in a pseudo-regular expression syntax. Hence, links such as “`http://www.example.com/foo/top.html`” would be included in the mirror by default. Of course, additional rules might be necessary depending on the site being mirrored: the default expression shall then be customizable. At last, duplicate links must not be transmitted to the crawler twice: the parser has to keep the state of all known URLs and avoid retaking multiple times links that were already taken.

The parser also has to handle very numerous syntaxes, which can mix relative or absolute URL forms, HTML escaping[6] (such as ` `), URL escaping[7] (such as `%3a`), and in general a loose syntax. This syntax tolerance

[4] See [2396], Sect 1.4 “Hierarchical URI and Relative Forms”
[5] See [1808].
[6] See [1866], Sect 14. “Proposed Entities.”
[7] See [1630].

from browsers is high: even with really broken pages (including errors in tag syntax), browser will generally do their bests to pursue its analysis, rendering "what could be understood."

As an example, the absolute link form:

"`http://www.example.com/page 2.html`" can be referenced using multiple syntaxes, including several incorrect or unadvised ones. In any ways, the URL has to be recognized and took in account as the browser would have done.

<a href= "page 2.html"> (*double quoted link)*

<a href= 'page 2.html'> (*single quoted link)*

<a href= "page 2.html"> (*HTML-escaped characters)*

`<a href= "page%202.html">` (*URL-escaped characters)*

`<a href= page%202.html>` (URL-escaped characters, no quote)

`<a href= page%202&#46;html>` (multiple-escaped characters, no quotes)

`<a href= "page 2` [CR]`.html">` (*unexpected carriage return)*

`<a href= "http:page 2.html">` (protocol in URL, but no host)

<a href= "//www.example.com/page 2.html"> *(no protocol scheme)*

<a href <a href= "page2 [CR] .html"> (*broken tag syntax)*

At last, links have to be rewritten to fit the mirrored website structure. Links using the absolute form, such as "`http://www.example.com/index.html`", needs to be converted into relative form, such as "`index.html`" ; and links beyond the mirror scope (links that did not match the default regular expression scope) have to be rewritten in their absolute form. Hence, mirrored pages needs to be modified to be useable in a local structure.

4.2.2 The Script Parser

Several months after the beginning of the HTTrack development, and despite of improvements in the HTML parser, there were a fairly numerous websites that were not correctly copied, with a lots of missing images, missing files, causing navigation errors, because the parser just did not "see" these links.

Inside HTML pages, specific scripting[8] zones must be considered, such as JavaScript (active code inserted in pages), which demand specific parsing. Unlike HTML tags, which are objects rather easy to analyse, script code is nearly impossible to fully handle: the logic behind variables, functions and expressions can potentially be unreachable. First, even with a

[8] See the ECMAScript scripting generalization, [ECMA-262].

complete JavaScript interpreter, actions triggered by mouse position, clicks on elements, or environment (the time, client variables and in general, environment entropy…) can not be captured. Second, the capture of links using an interpreter would not solve the other problem: modifying the code logic to "fit" the mirrored site. Detecting links is not sufficient: we also have to modify them. And if this is a fairly easy thing to do inside HTML tags, doing the same inside complex scripting code is nearly impossible.

Hopefully, in most cases, the JavaScript code used is simple enough to fit the limited analysis abilities of a program. To dynamically load images – or to cache them in the background, Web designers generally use direct assignment to object properties using static strings, such as "`foo.src=bar.gif`" or, to open new windows, use expressions such as "`window.open("foo.html")`." A rough 80% of links hidden inside script zones can be detected – and modified – by handling these simple cases. The remaining cases, using expressions or unknown methods, will just be left as it. The result will not be perfect, and – concerning HTTrack – we knew from the beginning that everything will not be. The objective was to reach an acceptable quality level, that would allow to take care of most sites. With similar algorithms, CSS (style sheets) zones can be parsed with an acceptable quality.

The following simplified automaton describes the text strings extraction inside scripting areas (such as <script> tag sections). After extraction, the string analyser attempts to guess whether the data appears to be a link or not, depending on its form: strings terminating with a known extension, such as "`.gif`" or "`.html`," or strings starting with a protocol part such as "`http:`."(see Fig. 4.2).

Additional heuristics are necessary to limit detection errors in the string analyser, and especially the characters preceding the string: substrings inside expressions must be avoided, such as in these two examples.

foo.src = "/images/welcome.gif";
bar.load("/docs/welcome.html");

We can assume that "/images/welcome.gif" and "/docs/welcome.html" refers to the current document's location – this simple heuristic is safe in most cases.

foo.src = dir + "/images/welcome.gif";

No assumption can be made for "/images/welcome.gif": the preceding character "+" clearly shows that the string is part of a string expression, and therefore the complete URL cannot be guessed without language analysis.

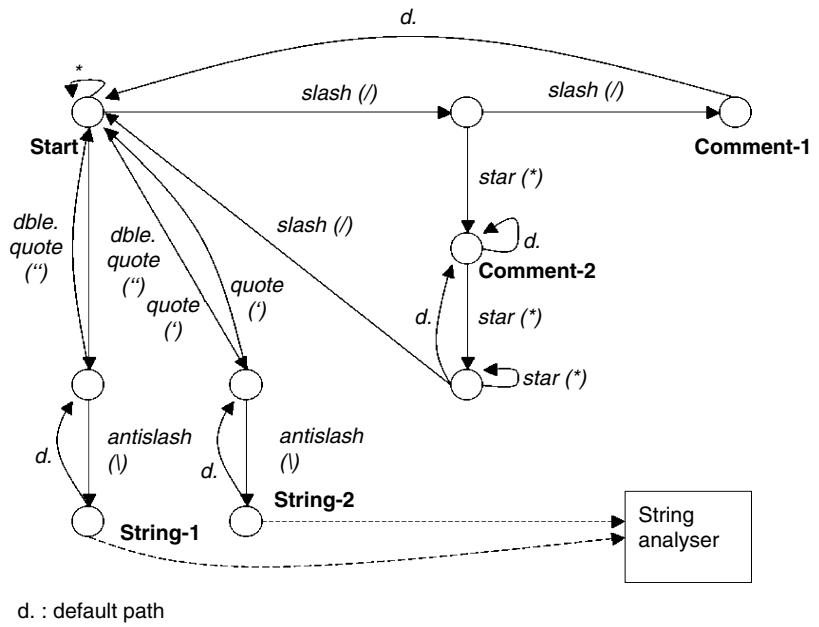


Fig. 4.2. Text strings extraction in scripting areas

Safe preceding characters includes “,”, “(“, “=”. Safe following characters includes “;”, “,” and “)”.

At last, strings selected as “probable URLs” must be modified to fit the local copy, as we do inside the HTML tags.

Having completed the automaton and the string analyser, we are able to handle most JavaScript cases. Greatly improving this algorithm is unfortunately very difficult, and would probably involve function and expression analysis, based on javascript language specification, and other advanced heuristics that are gar beyond the scope of a generic parser.

4.2.3 The Java Classes Parser

Another challenge is the handling of Java applets, which were quite popular at these times. After a while, the Microsoft-centric activeX format emerged, until waves of Trojans and viruses exploited this new format. Nowadays, the trend is more to use Flash applets, especially to design hideous and irritating advertisements you can not block. But in any cases, the annoyance for Web archivists remains exactly the same.

Let it be clear: if embedded scripting is annoying, binary embedded files such as java or Flash are a real pain in the neck. With text-based files format such as HTML or XML, you can easily modify subparts of the file without bothering about other ones. You can change a description in a far level; it will not impact the whole file. You can disregard most elements without even knowing their meaning. To summarize, you can focus on very specific elements – such as links – modify them, and forget the thousands other ones. Because these formats are intended to be simple, flexible, and easily understandable by human beings, and you do not need to understand the thousand-pages specification to get one single information. But binary formats such as Java classes rely on complex structures[9] that cannot be modified easily. Dare you modify a single string inside the file, and the whole file can be spoiled, because the string size, location and possibly content may have been referenced elsewhere using a collection of obscure pointers, offsets, checksums, and other magic tricks. For each binary format, you have to implement some complex algorithms to disassemble the desired data. Modifying these data and reassembling the whole thing is even more complicated.

A reasonable strategy to handle java classes is similar to the JavaScript heuristic, able to handle simple cases, not want a perfect binary Java parser. Hence, a basic class analysis based on embedded strings inside the .class file data segment is sufficient to enumerate interesting data (e.g., URLs). Here again, strings that "looks like" URLs will be kept. But the main difference is that we will not make any changes inside the binary file due to the complexity of this task. Besides, embedded strings representing relative links would probably work on a local environment, but not fully-qualified ones: modifying the strings without understanding (or analyzing) the logics behind would not be sufficient anyway.

Scanning interesting strings inside a Java .class file is not very complicated: we first have to (down)load the class file header in memory (10 bytes) and check the 32-bit integer magic ("0xcafebabe") to ensure that the file is explicitly a java class. See below the .class file header. The constants pool includes all static strings, such as strings used for URLs, and indexes of class name strings used to include specific java classes.

```
ClassFile {
    u4 magic;
    u2 minor_version;
    u2 major_version;
    u2 constant_pool_count;
    cp_info constant_pool[constant_pool_count-1];
```

[9] See "The Java Virtual Machine Specification – The class File Format."

```
        u2 access_flags;
        u2 this_class;
        u2 super_class;
        u2 interfaces_count;
        u2 interfaces[interfaces_count];
        u2 fields_count;
        field_info fields[fields_count];
        u2 methods_count;
        method_info methods[methods_count];
        u2 attributes_count;
        attribute_info attributes[attributes_count];
    }
```

Then, continuing to following the JVM Specifications,[10] enumerate all constant objects (`constant_pool` structures), and analyse all strings (`CONSTANT_String` and `CONSTANT_Utf8` objects) inside the file. Imported java class names can be detected using their corresponding objects (`CONSTANT_Class_info` structures): all specific (i.e., not standard library classes located in the java/ section) imported libraries can be retrieved later, as embedded links inside HTML data would be retrieved.

See Table 4.1: the "interesting" objects to scan inside a .class file are static strings that can potentially represent URLs, and Classes string indexes used to import external libraries (classes).

All detected links are then transmitted to the crawler, and the java class is stored untouched.

Despite a relatively high number of bogus mirrors, the result is rather positive: reasonably simple applets generally work, including applets composed

Table 4.1. Objects to scan inside a class file

Constant type	Value
CONSTANT_Class	7
CONSTANT_Fieldref	9
CONSTANT_Methodref	10
CONSTANT_InterfaceMethodref	11
CONSTANT_String	8
CONSTANT_Integer	3
CONSTANT_Float	4
CONSTANT_Long	5
CONSTANT_Double	6
CONSTANT_NameAndType	12
CONSTANT_Utf8	1

[10] See "The Java Virtual Machine Specification – The class File Format."

of multiple classes. But more complex applications generally fails due to missing files; either because the link was not modified inside the class, or because the link was invisible to the parser due to the hidden logics behind the binary code.

4.3 Fetching Document

The other engine element in a copying tool architecture is the robot responsible for gathering data from online websites – HTML pages, images, style sheets, and in general any media file available on the server. A stack of URLs to be collected, initially filled with one or more "root" addresses of HTML pages given by the user, is used by the robot which connects to respective servers, sending requests, and handling downloads. This robot can work in parallel of the parser to improve performances: while the parser is scanning pages, the crawler downloads data using multiple connections, dispatching ready files to the parser.

The crawler/parser interactions can be summarized in the following diagram: these two processes share a common bucket of links – the heap – filled by the parser as new links are being discovered, and emptied by the crawler as new links are being successfully downloaded (see Fig. 4.3). You can see it like a perforated bath tub: the number of remaining links to be downloaded varies with the time, to reach the count zero when the website is completed. At this point, the mirror is considered finished. Note that the parser only scans files known to contain links: HTML pages and other limited formats (such as Java or Flash files). Other files (such as images) will be stored "as is" on disk, without modifying them.

To fetch files, the crawler first fetch an URL in the link heap, and decompose it in three components: the protocol (http, https, etc.), the authority (the Web server hostname), and the path (including the query string – the optional part before the "?"). Note that the fragment (the optional part after the "#" character) is not part of the URI, and therefore never included when collecting URLs.

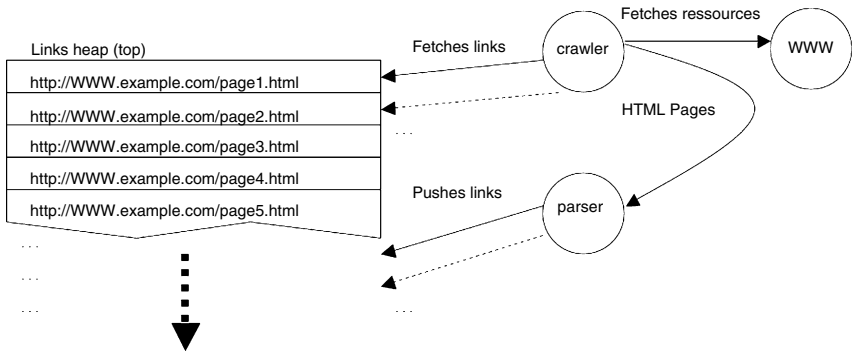


Fig 4.3. Crawler/parser interactions

The protocol will be used to dispatch the main data collecting routine: HTTP or HTTPS (HTTP with an additional SSL[11] layer), optionally ftp[12] or even the file[13] pseudoprotocol. Most links in the WWW will use the HTTP protocol, which is the only part we will discuss here.

When fetching an HTTP resource, four steps will potentially consume time or create latency:

- Hostname DNS resolution (wait for the DNS server response);
- Connection to the website (three-steps TCP connection handshake);
- Request sending (not really significant: less than half a Kilobyte, that is, one TCP packet – the ping latency will be more important here);
- Response fetching (depending on the transfer rate).

These steps can be processed sequentially for all links to be fetched. Or they can be optimized for most of them:

DNS resolution can be cached in the crawler so that further requests can be fulfilled without delays

Network latency created by connections establishments can be greatly reduced by using HTTP 1.1 Keep-Alive connections and a pool of connected sockets managed by the crawler

Response transfer rate can be boosted by using HTTP compression, especially for HTML data

[11] See [2246].
[12] See [959].
[13] See [1738], Sect. 3.10.

At last, specific HTTP handling such as the ability to continue the download of an interrupted file, can improve the overall performances on sites suffering from network instabilities

4.3.1 Authentication, Session, and Endless Loops

Many websites are not friendly to crawlers (and generally, nor to search engines) either because their hypertext pages uses technologies hostile to crawlers, such as Java/JavaScript or Flash, or because of their internal navigation scheme. Authentication or sessions are a first difficulty. Session-driven websites generally attach a "ticket" to each visitor when accessing the homepage, under the form of a cookie, an opaque data transmitted by the server to be used consecutively by the client for further HTTP requests. Authentication schemes either use username and password credentials, cookies, or both of them. Failing to pass the expected ticket can prevent you from accessing pages beyond the main entrance. Hence, the handling of cookies[14] or authentication[15] is not an option, nor the secure HTTP version, HTTPS,[16] which is the preferred way of securing the authentication process by avoiding any clear text credentials exchange. But cookies can also be replaced by identification elements included in all URLs as query-string parameters, or directly as part of the path segment. Entering of credentials can be expected through forms, not well suited for automated crawlers.

Each of these difficulties are annoying issue, as the website will probably attribute different identification elements each time the site is being updated, possibly rendering update impossible. Certain session-based navigation systems may also reattribute different tickets time to time after an arbitrary expiration time was reached, causing the robot to loose its way inside the site, and pushing it in endless loops and infinite mirroring. Other sites, especially those hosting forums or message boards, are using multiple different URL syntaxes to reference a unique page: duplicate files for each messages tend to dramatically increase the overall size being downloaded.

An attentive analysis of such sites, use of specific scan rules, and multiple tries will generally allow to overstep these problems. But it means manual, nonautomatic adaptations: fully automatic copies are sometimes beyond the capacity of mirroring tools.

---

[14] See "Persistent client state HTTP cookies" and [2965].
[15] See [2617].
[16] See [2818].

4.3.2 Redirections, Refreshes, and Frames

Both HTTP protocol and HTML specifications offers different ways to redirect a browser to another location, automatically: the browser first goes to the desired location, and is then automatically redirected to the new address without any user interaction. The first – and cleanest – method is to use a specific HTTP response (HTTP error codes 301, 302, 307), which will not return the final document, but instead it is new location. Websites moved to another place will still be accessible using their former address; which will gently redirect to the new one. This feature can also be implemented using specific HTML tags, metatags, that can replace certain HTTP headers such as the media type (and, for HTML pages, the character encoding being used) or the status code itself, using the "refresh" feature. At last, invisible frames can be used to "embed" pages located in another place. All these tricks are generally invisible to the user. But not to crawlers, that needs to take a decision: must the various redirections be followed, or not? Following by default redirections can be useful, if the site was moved elsewhere. But it can also be dangerous, it the site uses this technique to generate a list of external links that can be logged by the server (for popularity statistics), or if the site uses redirects to the main page instead of "404 errors" – a bad habit used by many free hosting providers. For this reason, the website copier HTTrack is not following by default redirects – except if the redirected place is allowed by the scan rules given by the user.

4.3.3 Connectivity

When used in corporate environments, Web crawlers face network restriction problems, such as the inability to access the Internet directly. Handling of proxies (delegating server used to fetch remote resources) is then an appreciated feature. Proxies are also used as protection and to accelerate Web responses inside internal networks, allowing to get a fast access to pages already visited. This interesting feature was for example exploited by administrators who used HTTrack to pre-cache external websites before presentations, providing better response times during the show as the proxy already had all the pages in cache! HTTPS is also a nice feature to be implemented for corporate environments, and not only for authentication issues: many corporate sites are only accessible through a secured connection. At last, implementing the next-generation protocols such as ipV6 is not an option: despite its currently limited usage range, it should be progressively widely used.

4.3.4 Politeness: Bandwidth and CPU Limits

Implementing bandwidth limiters and limiting the number of connections is clearly an important security feature for a WWW crawler. This protection is specially important with nowadays lines, as Internet Providers can now offer several megabits of bandwidth in urban areas even for domestic lines. A regular user can potentially cause denial of services when downloading a website at "full speed." Abusing the bandwidth would not only cause legitimate complaints – but would also lead to ban mirroring tools everywhere, ruining all archivists efforts.

As an example, even in the beginning of the HTTrack's development, bandwidth and CPU usage were important issues. The bandwidth available for testing and validating the crawler was fairly high in our development environment, allowing to run multiple crawls in parallel or reissue regular benchmarks to fix bugs, detect new problems, and improve the engine in general. It was rather easy to get very fast transfer rates – too fast for many servers, which could be quickly overloaded.

This remark is also true when accessing resources that require processing time on the remote server, notably when fetching dynamically generated pages with underlying costly processes such as database operations. Using multiple connections to access these pages will lead to overload the machine serving the data, causing another kind of denial of service.

4.3.4.1 Politeness: robots.txt

In many cases, websites containing area crawlers should not automatically process: large files, costly dynamic pages, sections with potentially endless loops, and in general parts that are not suitable for fully automated processes.

For these reasons, webmasters now places specific hints for robots that can be found on the server in a file called "robots.txt." This simple ASCII file lists areas that crawlers should or should not visit for various reasons, using a simple syntax, optionally specifying names of crawlers concerned by each rules. The robots.txt[17] standard is an important feature to implement when designing search-oriented crawlers, but in certain cases, archiving crawlers will have to bypass them when areas not suitable for indexing – but suitable for archiving – have to be processed. In such cases, Webmaster's cooperation is highly recommended, especially to suggest which areas can be safely crawled.

[17] http://www.robotstxt.org/wc/exclusion.html

Example of robots.txt rule explicitly forbidding the/private subarea from crawling:

*User-agent: **
Disallow: /private

4.4 Create an Autonomous, Navigable Copy

From the beginning, an implicit choice was made for the crawler: as a regular (human) user would, copied files are stored in a filesystem (i.e., on a disk as regular files) rather than inside a database, for example. The result would be directly viewable with any browser, as if you were accessible the "real" website. This choice is not only an intuitive choice. It is also a security choice. A database-driven copy would require the corresponding database software to be useable, not necessarily compatible with all operating systems and architectures. This software will possibly be deprecated in several years, and impossible to run on future systems. An archive that might be unusable in the future because of a software component would be a grave design choice for preservation purpose. Similarly, any nonstandard software that might be necessary to view the copied website would fall in the same critical problems. Copying Web resources into regular files that would be easy to backup, mapping their names as regular filenames, and allowing any standard (i.e., respecting Internet standards) browser to access them is an obvious and reasonable choice. Being independent from any application vendor, operating system vendor and even machine architectures is one of the keys of the Web preservation.

Another options to use standard container files like WARC but at the time we started developing HTTrack, this standard did not exist.

Copying live sources such as HTTP files to local file systems and browsing them without the need of any additional programs but a regular browser has several drawbacks. The first problem lies in the way you organize the file structure locally. Online resources rely on URLs which, unlike filesystem paths, follow arbitrary naming convention that depend on the remote server implementation. A Unix fully qualified path "/home/users/smith/document.tex," or a Windows fully qualified path "C:\Documents and Settings\smith\My Documents\document.doc" both follow strict conventions, such as characters allowed in filenames and directory names, restrictions on their size, and a specific character (/or\) playing the role of a separator. In the contrary, Web Servers can choose to follow a traditional structure, such as

`http://www.example.com/foo/bar.html`, or in the contrary use exotic naming conventions, such as `http://www.example.com/foo; bar; t=html;q=1//`. Luckily, most servers are gentle enough to avoid such eccentricity – hence we can consider that in most cases, URLs will look more or less like regular paths, allowing to recreate a similar filesystem structure when mirroring a site. The URL "`http://www.example.com/foo/bar.html`" will then be copied as "bar.html" somewhere in a subdirectory "/foo/bar/." However, many links do not have any document name, and use the default "/" convention. The link "http://www.example.com/foo/bar/" can be stored in the "/foo/bar" subdirectory – but it also needs a filename, such as "index.html" or "default.html." One of these arbitrary names will then be choosen. But what to do if both "http://www.example.com/foo/bar/" and "http://www.example.com/foo/ bar/index.html" exists in the website we are copying? Similarly, a number of forbidden characters must be avoided because they are forbidden or discouraged on Windows[18]/Unix[19] environment, and additional restrictions must be taken in account if the website is to be stored on CDRom media[20]. At last, Windows filesystems can handle Unicode characters in filenames such as accents; characters that can not always be represented easily on all platforms. In all these cases, the offending characters can be changed by replacement characters such as "_" (underscore), and names shall be truncated if necessary – but other strategies could be used, such as escaping these characters using an arbitrary convention like URL-escaping.

Another scenario has to be handled, too. When opening a document on a local disc, the file extension, such as "pdf" or "html" also describes its type: an Acrobat document, or an HTML file. If you rename an HTML document as "pdf," you will not be able to open it anymore: the wrong program will be launched to view it, and will fail to do so. On a Web environment, media types are transmitted by the server to the client through HTTP headers: the URL naming scheme is not important anymore. The resource "`http://www.example.com/index_1.cgi`" can be an HTML document, a PDF document, an image, etc. it is the media type transmitted by the server when fetching the document that will be decisive. We have to rename the resulting document if we want it to be useable in a non-Web environment: index_1.html, index_1.pdf or index_1.jpg depending on the media given by the server.

---

[18] The characters / : * ? " < > | are forbidden, among with other restrictions (case insensitivity , reserved names…).

[19] The use of ~ | * is discouraged for shell-expansion reasons.

[20] See ISO9660 restrictions, notably filename size limits (30 characters) and forbidden characters * / : ; ? \.

Here again, the consequences of these adjustments have to be considered, as different URLs can have identical local names after applying naming restrictions. Consider the following four URLs pointing to an HTML document:

http://www.example.com/index_1.html
http://www.example.com/INDEX_1.HTML
http://www.example.com/index:1.html
http://www.example.com/index_1.cgi

All these links would get the same filename, "index_1.html," as the second URL is identical to the first one except for the case (but Windows environment are case insensitive), the third one contains the unadvised character ":," replaced by "_," and the fourth one will be renamed "html" instead of "cgi" to be viewable offline in a filesystem. Avoiding such collisions means that different names must be found when they occur – like "`index_1.html`," "`index_1-2.html`," and "`index_1-3.html`," and require adjustment in the parser which task will also be to apply these changes in downloaded HTML pages so that links can refer to the respective local files.

4.5 Handling Updates

4.5.1 Updating the Copy

One drawback in making static copies of websites is that these copies will not change anymore, they will not be updated by their original authors, and as books in libraries hold their typing errors and mistakes, copies remains exactly the same as they were during their creation. They may deprecate. Depending on the archivist's needs, it might be desirable to make regular copies of preserved websites either to and get up-to-date version, or to regularly store a copy that would allow to retrieve the site on a specific moment.

One solution is to retrieve the entire website when necessary, repeating exactly the same operations. But when making regular copies of websites, updating (i.e., only transferring modified content) tend to be a major issue, especially with big websites containing or with large media files. Recrawling exhaustively a website in the purpose of having an up-to-date copy is a very inefficient method: waste of bandwidth, waste of time, and waste of storage space when handling multiple site versions. A regular browser usually stores recently pages and associated files in a specific location

called “cache,” which can be used to avoid retransferring data that was previously consulted. When visiting pages already in cache, the browser is able to ask the remote server whether its local copy is fresh or if it has to be transferred again. Here again, the solution for updating problems is to mimic browsers, by handling a cache that will be used by the crawler to check the freshness of already downloaded data.

There are two main mechanisms described in HTTP 1.0[21] and HTTP 1.1.[22] The first and most widely used one is the old HTTP update mechanism, which rely on the remote document date to ensure that the resource is always up-to-date. By sending the document’s date to the client, the server allows it to perform further update check by asking something like “Please give me the /index.html document, except if it was modified since 14 July 2002.” The second mechanism is a more general system, which uses an opaque string aimed to identify a specific resource content. It can be the document’s md5 checksum, for example, or any other element that can be used by the server to identify the document’s version/freshness: it can also be the document’s last-modified date, but this is only a particular case.

The diagram above gives a rough idea of a caching mechanism (see Fig. 4.4). When fetching a link, the crawler first checks whether it is already known by the local cache. If not, the file is downloaded as usual. But if the cache already has a version, the crawler can either decide to ask the server for freshness, or to directly take the existing cached file – when recovering an interrupted or crashed or mirror session, for example.

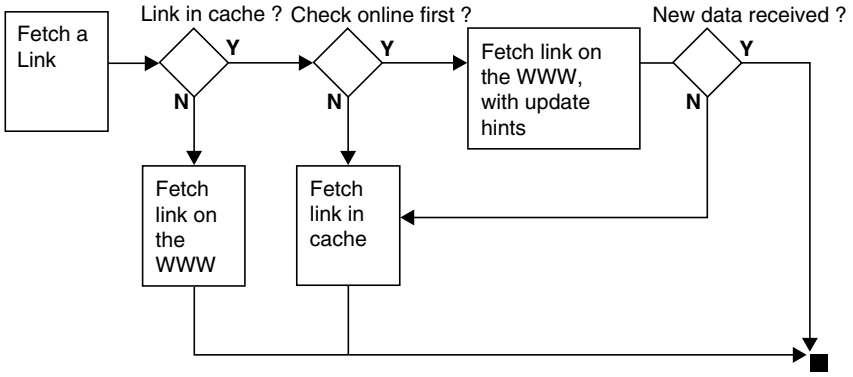


Fig. 4.4. Caching mechanism

[21] See [1945], Sect. 10.9.
[22] See [2616], Sect. 14.19.

The cache can store files such as images or binary data, or reference their location if they already exists on the mirror file tree: in this case, original HTML data needs to be stored "as is" anyway in the cache, because the existing files which were modified by the parser and can no longer be used effectively due to URL mangling (you can not guess the original URL because of that). Additionally, Web server metadata must be stored among file data such as original media type, status code, and of course information needed for updating purpose: the "Last-Modified" date, and/or the "Etag" opaque string. Note that the cache files are not necessary to view the mirrored website, but only to update it, and thus can be omitted when making navigable copies such as CD-Rom extractions.

4.5.2 Storing the Updated Copy

Being able to update a copy is one thing. Propagating the update in the copied archive is another one. On a filesystem tree, updated files can replace previous versions (overwriting them) and new HTTP header information can be merged in the cache. But updated files can also be handled using a file system managing versioning, such as a CVS[23] tree, allowing to crawl dated copies without needing to store multiple copies of the same files. If control version directories are too difficult to implement because of the necessary versioning interface, or are not desirable because is ties up the archive to a specific software, an intermediate solution is to use native file system's linking (on Unix systems, symbolic or hard links) and organize versions in different directory trees, with files or directories either pointing to the preceding copy, or storing the new version depending on the website changes.

On this example, the first copy made in January is 50 MB large. The update, issued in February, modified an overall of 3 MB of data. The remaining files (47 MB) were untouched since January. The third run modified again 2 MB of data, and brought 2 MB of new files (4 MB of new material overall), with 46 MB of remaining files untouched since January, and 2 MB of files untouched since the February update. In this imaginary example, the three sites versions represent a total of 152 MB of virtual data, for a total of 57 MB of physical data (see Fig. 4.5).

In many cases, updates are far less expensive than the "first run": handling versioning is generally an inexpensive feature compared to multiple identical copies.

[23] Control Version System.

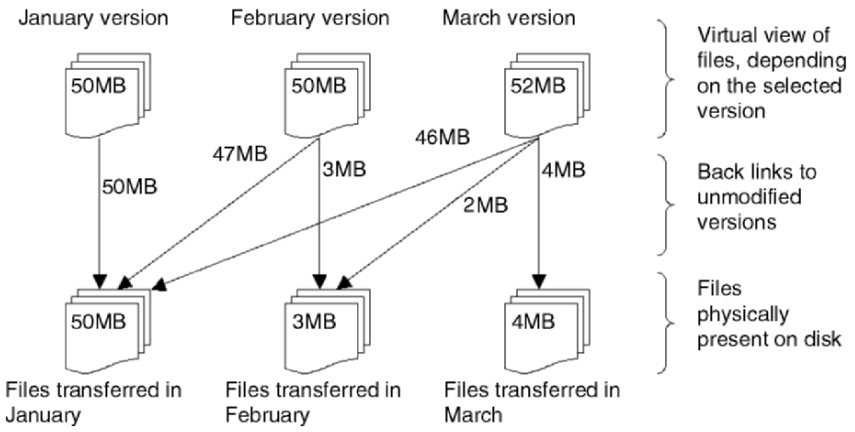


Fig. 4.5. A website copy with two updates

4.6 Conclusion

Preserving and archiving websites is a thrilling technical challenge which must be pursued, endlessly. Evolving technologies, evolving contents and evolving growth of the WWW means that no definitive archiving solution might ever be found. Existing solutions can still be improved as no perfect system was yet made. This continuous work in improving Web preservation techniques is done by passionate people around the world, in libraries, universities, companies and by individuals.

Reference

The WWW was built over Internet standards, especially through the Request For Comments (RFC), which describes most protocols and standards when using Internet technologies. They are public, free, and easily understandable for "regular computer programmers," unlike many other international standards. This easy access was one of the reasons why the entire Internet grew so quickly: a common standardized technical base that everybody (i.e., not only accredited companies) could use.

Below you will find several RFC highly recommended when developing WWW preservation tools. This list is not exhaustive, and shall be completed by references indicated at the end of the documents described below.

HTTP References:

– RFC 1945 – Hypertext Transfer Protocol – HTTP/1.0
(first version of the HTTP protocol)
http://www.ietf.org/rfc/rfc1945.txt?number=1945
– RFC 2616 – Hypertext Transfer Protocol – HTTP/1.1
(second version of the HTTP protocol)
http://www.ietf.org/rfc/rfc2616.txt?number=2616
– RFC 2818 – HTTP Over TLS
(also called "https")
http://www.ietf.org/rfc/rfc2818.txt?number=2818
– RFC 2660 – The Secure HyperText Transfer Protocol
(not to be mixed with https)
http://www.ietf.org/rfc/rfc2660.txt?number=2660
– PERSISTENT CLIENT STATE HTTP COOKIES
(browser "cookies")
http://wp.netscape.com/newsref/std/cookie_spec.html
– RFC 2965 – HTTP State Management Mechanism
http://www.ietf.org/rfc/rfc2965.txt?number=2965
– RFC 2617 – HTTP Authentication: Basic and Digest Access
Authentication (used when authenticating to a server expecting credentials)
http://www.ietf.org/rfc/rfc2617.txt?number=2617

URL/URI References:

– RFC 1630 – Universal Resource Identifiers in WWW
http://www.ietf.org/rfc/rfc1630.txt?number=1630
– RFC 1738 – Uniform Resource Locators (URL)
http://www.ietf.org/rfc/rfc1738.txt?number=1738
– RFC 1808 – Relative Uniform Resource Locators
http://www.ietf.org/rfc/rfc1808.txt?number=1808
– RFC 2396 – Uniform Resource Identifiers (URI): Generic Syntax
http://www.ietf.org/rfc/rfc2396.txt?number=2396

HTML, Java and Script References:

– RFC 1866 – Hypertext Markup Language – 2.0
http://www.ietf.org/rfc/rfc1866.txt?number=1866

– Standard ECMA-262 - ECMAScript Language Specification
(standard based on JavaScript and JScript)
http://www.ecma-international.org/publications/standards/Ecma262.htm
– The Java Virtual Machine Specification – The class File Format
(binary specification of class files)
– http://java.sun.com/docs/books/vmspec/2ndedition/html/ClassFile.doc.html
– The Java Language Specification
http://java.sun.com/docs/books/jls/second_edition/html/jTOC.doc.html

Other Internet Protocols References:

– RFC 2076 – Common Internet Message Headers
(notably headers used in HTTP headers)
http://www.ietf.org/rfc/rfc2076.txt?number=2076
– RFC 2822 – Internet Message Format
(base RFC for many Internet protocols)
http://www.ietf.org/rfc/rfc2822.txt?number=2822
– RFC 2045 – Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies
http://www.ietf.org/rfc/rfc2045.txt?number=2045
– RFC 1950, RFC 1951, RFC 1952 – Compressed Data Formats
(used in HTTP compression)
http://www.ietf.org/rfc/rfc1950.txt?number=1950
http://www.ietf.org/rfc/rfc1952.txt?number=1952
http://www.ietf.org/rfc/rfc1951.txt?number=1951
– RFC 2246 – The TLS Protocol (used in https)
http://www.ietf.org/rfc/rfc2246.txt?number=2246
– RFC 959 – FILE TRANSFER PROTOCOL (FTP)
http://www.ietf.org/rfc/rfc959.txt?number=959

5 Archiving the Hidden Web

Julien Masanès

European Web Archive
julien@iwaw.net

5.1 Introduction

As we have seen in the earlier chapters, the main method for gathering Web content is based on path finding. This is due to the fact that the HTTP protocol does not provide a full listing functionality. Therefore, each page must be "discovered" by crawlers in the first place in order to be fetched and archived. We have seen in Chap. 1 that this introduces an important time constraint on the whole gathering process. But it also entails that, at least one path is found for each document in order to be archived. This is far from being always the case. Actually a large portion of the Web cannot be reached by automatic tool for this very reason. This portion has been first call the invisible Web in 1994 by Jill H. Ellsworth (Bergman 2001) as the fraction of the Web that is not indexed by search engine's crawlers. It has later been proposed to call it the "deep Web" in opposition to the "surface Web" or the publicly indexable Web (PIW) (Lawrence and Giles 1999) that crawlers can easily access to.

We would not use both the terms in this book because they carry too much ambiguity. The first one, "invisible Web," could lead one to think that the problem is a problem of displaying or rendering pages whereas it is a problem of accessibility for automatic tools. The second one, "deep Web," can be confused with the depth of document in the hierarchical hypertext structure of websites. We will use instead "hidden Web" to designate the portion of the Web that crawlers cannot reach, bearing in mind that, first this is a technically defined limit, which implies no specific type of content nor common human experience of navigation's limits whatsoever. And secondly, this definition being purely technical, the frontier it sets out may change as technology evolves.

For instance, sites built with traversal functions coded in Flash used to be considered as hidden Web before Macromedia released an SDK that

enables extraction of links from flash code, which makes it possible for crawlers to find their path through flash coded websites. This kind of website can therefore not be considered as hidden Web any longer.

Although most of the attention will be drawn in this chapter as well as on the rest of the literature, on crawlers limitations, it is worth noting that crawlers also tend to discover pages that are rarely if not never seen by any human. Actually one of the few study on this issue by Boufkhad and Viennot (2003), has shown, using Web servers logs of INRIA for 2002, that users "missed" a comparable number of pages that crawlers do.

5.2 Finding At Least One Path to Documents

Connected pages of a website form together a Web of paths to access documents. Any single page can be linked from one or several others and crawlers will reach it from the first one it gets. Pages non-linked from the rest of the website are not part of it from this point of view, even if they dwell on the same server. This Web of path can be extended to any document that can be generated or access through virtual path constructed via a query. Dynamic pages for instance, generated out of content stored in database, exist only virtually but can be considered as part of this Web of path. The important point here is how path to objects are created and hence whether crawlers can find them or not.

Two general cases can be distinguished here: in the first one, there is a limited and predefined sets of path value. This is the case of hyperlinks and menus for instance. Crawlers can extract, interpret and follow such path depending on the language used to encode them.

The second case, the most difficult for crawlers, requires an explicit user interaction (i.e., more than one click), hard for crawlers to mimic. This is for instance the case when users have to enter query terms via HTML forms to get a specific document like an image, an article or a page dynamically generated. This type of hidden Web, that we call structural hidden Web, holds a very large amount of content as it has been used to publish on the Web large document repository, both structured (database) and unstructured, i.e., collections (images, scientific papers, music, etc.) (Storey and Jahnke 1999). A convenient way of publishing such pre-existing large collection of content is indeed to offer access dynamically through databasedriven gateway containing descriptive information of each item, instead of hard-coded links. This typical information architecture (form, database and collection) is used today by many sites, which can be considered as part of the hidden Web. We call this type of architecture

documentary gateway as it provides an entrance to large information space through an explicit interaction (search) (Masanès 2002). In some cases, an alternative-browsing interface is provided so that crawler can still access content of such repository in which case they fall in the first type.

We will now review each type of user interaction in more details and see which problem they raise for crawlers.

5.2.1 Type of User Interaction

Ludäscher and Gupta (1999) proposed a model of interactive source. It comprises four types of input elements:

- Hyperlinks are the classical way to provide user input limited to one possible traversal path;
- Menus are used to select a subset of values from a predefined set of traversal path;
- Forms can be conceived as dynamic links with multiple input attributes ranging over an infinite domain;
- The fourth type of input element can only be according to this model, supported with explicit user interaction. They are called non-wrappable elements (image maps, GUIs based on java etc.).

For a real-world example of distribution of various input elements in medical images-rich websites see Frankewitsch and Prokosch (2001). Each of these types of input elements is a different case for crawlers. The first two (hyperlinks and menu) raise only interpretation problems for crawlers and form together what we will call defined-value input elements.

Open-value input elements (mainly HTML forms), in opposition to the latter, raise the issue of reducing the space of possibility to one or several defined paths. They are used in many different contexts and we will discuss them further below. The fourth type although less common than forms, is even harder to access for crawlers. For a detailed typology of problems encountered by crawlers see the two following IIPC reports (Boyko 2004; Marill et al. 2004). We will outline in the following section the main cases and discuss what problems they raise for crawlers.

5.2.1.1 Defined Value Input Elements

Plain HTML links is the most common of this type, enabling easy path finding. Relative links may however raise some issue to interpret. They are based on rules described in RFC 1808 and RFC 2396 and are similar to Unix addressing. Relative location starts from the current directory and go

down using slashes (/) and up through parent directory using two dots (..). A relative path starting with slash means the root directory of the host. The difficulties come from bad usage by page creator or content management system. This includes for instance extra-dots, supernumerary parent directories or other bizarre combination. If present, the base tag in the head section of the HTML page provides an alternate default location for relative links from which all the links should start rather than the current directory. Java script links as well as other scripting language are widely used to generate links specifically for menus and scrolling navigation, dynamic calendars, etc. As a link can result of any combination of command, variable and/or user interaction input, it is almost impossible in some cases to interpret the link without executing the script.

Crawlers can have either rule-based interpretation[1] or try any possible combination of path and file names they find in the script.[2] In each case a complete success is not guaranteed. An alternative solution that has not been tested yet would be to interpret by executing the code instead of parsing it. This would mean to run browsers and simulate context and user interaction. Some other type of links like frames links, image links may also raise some difficulties to interpret but can generally be successfully followed.

5.2.1.2 Open-Value Input Elements

In all the previous cases, however, despite the problem encountered, a finite domain of possible paths is defined by the code. This makes a huge difference with the second type where an infinite and undefined set of values can be assigned to links.

The main mechanism for this is HTML form. They enable users to pass an arbitrary value to the server. Input values can be used for instance to query documents, resulting in the generation of a set of links to pages or documents. Entering a query containing an author and a title in a catalogue form can for instance result in a list of publication, with links to each of them. These links have been generated from a database and embedded in an HTML result page. If there is no other link pointing to these documents, for instance an alphabetically ordered list of publications, then crawlers will only be able to access them by the virtual path generated through the query interface.

---

[1] This is the case for instance of the open-source web copier HTTrack (see Chap. 3 (Roche 2006).

[2] Heritrix, the open source archive-quality crawler developed jointly by the Internet Archive and the nordic libraries in the IIPC, implements this approach (Mohr et al. 2004)

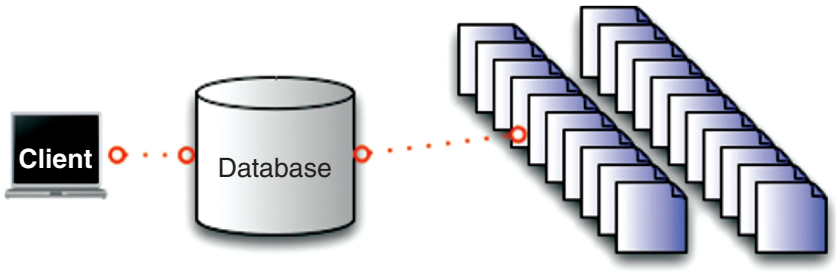


Fig. 5.1. Documentary gateways where access to large series of documents is made by querying a database

This type of architecture of information called documentary gateway (Masanès 2002) is very common and has to be distinguished from other use of forms (see Fig. 5.1). Forms are mainly used to gather user inputs like login, input of contact info or feedback, submission to forums, generic search box, etc. Cope et al. (2003) found that about 50% of HTML forms were search interfaces and (Lage et al. 2002) found in their sample that 95% of forms, including generic search box, were undesirable forms.

But even if most of forms found on the Web are used for other purposes, the remaining ones are the entry point to the huge information space that the hidden Web represents. Two studies have tried to characterize it.

5.3 Characterizing the Hidden Web

The first one, conducted by Bergman in 2000 (Bergman 2001), used overlapped analyses between pairs of search engines to estimate the number of hidden websites. They evaluated that there is a range of 43,000–96,000 hidden websites based on the presence of forms. Unfortunately, the filters used were not documented; it is therefore difficult to judge the value of these results and to compare them with others. They also made micro-analyses of the 60 largest ones and estimated the size of the hidden to be 550 billion pages, which is 550 times larger than the surface Web of that time.

The main assumption behind this study is that one item in a database corresponds to one page generated and its size is estimated “HTML included”.

This basis includes all HTML and related code information plus standard text content, exclusive of embedded images and standard HTTP “header”

information. Use of this standard convention allows apples-to-apples size comparisons between the surface and deep Web. (Bergman 2001)

For example, an entry in the national climatic database of the US (the largest example in their sample) would correspond to one page of 13 Kb.

The fact that this database and the NASA EOSDIS database represents almost 80% of the total size of the sample, shows that this study provides a biased view of the hidden Web toward repetitive and non-documentary content. However, this study draws for the first time attention on the importance and richness of this part of the Web.

A more recent study by Chang et al. (2004) provides a more detailed and documented characterization of the hidden Web. It distinguishes between structured database i.e., database associating key pair values and unstructured content (text, image, audio, video) that forms the content of what we call documentary gateway in this book. In Chang et al. study, as in most of the studies on hidden Web, main attention goes to the first type (structured database) as they are made by researchers from the database community interested in data integration from the Web. Nevertheless many findings are of interest from a more content oriented perspective, like ours.

They conducted a micro- and macro-studies. The macro-study was based on one million randomly generated IP addresses that were tested to find HTTP servers. Two thousand two hundred and fifty six were found and crawled and 126 were identified as hidden websites containing 190 Web databases. Reported to the global Web, this means that 307,000 sites contain 102,000 documentary gateways and 348,000 structured databases. This illustrates the importance of hidden websites. Chang et al. also identified the distribution of Web database over depth showing that 91.6% of them were found within depth 3.

In their micro-study they investigated 441 sources in more details. They first show that in many cases there are alternative navigational path to reach the content, in which case content is not really hidden for search engines. Actually "hiddeness" of websites depends on domain.

They studied query interface schemas and more precisely the number of attributes they comprise: the smallest size of schema is 1, the largest 18 and the average 6. They also studied the schema vocabularies and showed that the top five schema attributes are "titles," "keyword," "price," "make" and "artist".

As Chang et al. state it, this is encouraging for automatic processing can exploit such regularities:

> Our survey apparently indicates dual phenomena that together uniquely characterize the deep-Web frontier: First, as a challenge: Sources online

are virtually unlimited; even for a specific domain of interest, there is often an overwhelming number of alternative sources (the proliferating sources phenomenon). Thus, large-scale integration is a real challenge. Second, as an opportunity: However, while sources proliferate, in aggregate, their complexity tends to be "concerted", revealing some underlying "structure". In particular, we observe such concerted structure on the attribute vocabularies and query patterns across Web sources. Such aggregate vocabularies are clustering in localities and converging in sizes (Chang et al. 2004).

5.4 Client Side Hidden Web Archiving

As shown in the earlier section, hidden Web entrance point (form) exhibit regularities that have been used to try to automatically extract content with wrappers, sometime called hidden Web agents (Raghavan and Garcia-Molina 2001; Lage et al. 2002), see also Hearst (1998) and Adams (2001) for a general introduction on this topic.

The role of these agents is to detect HTML forms, learn to fill them in, identify and fetch resulting content. This can be implemented for providing a integrated search interface (see for instance, a review of this topic, Florescu et al. 1998, an example of search service implementation, Bergman 2001, and a recent prototype development in this domain, He et al. 2005). Detection is usually made during normal crawls by analysing pages that contain HTML forms. Heuristic are used to eliminate undesired forms (login page or contact information page, generic search box, etc.), then agents extract query fields and labels, try to compare these with known labels and sometimes probe them with known vocabularies to assess thematic (Gravano et al. 2003).

Finally, forms are automatically filled and resulting pages or documents are stored. Heuristics to filter out undesired forms used by Lage et al. (2002) allows to eliminate forms with less than a certain number of elements and forms comprising any password HTML type element. Cope et al. (2003) propose a more elaborated decision tree to discover search interfaces on the Web (see Fig. 5.2.) based on decomposition of HTML forms in atomic features like "http," "domain," text control, password control, etc. and is built from a training set and learning algorithm.

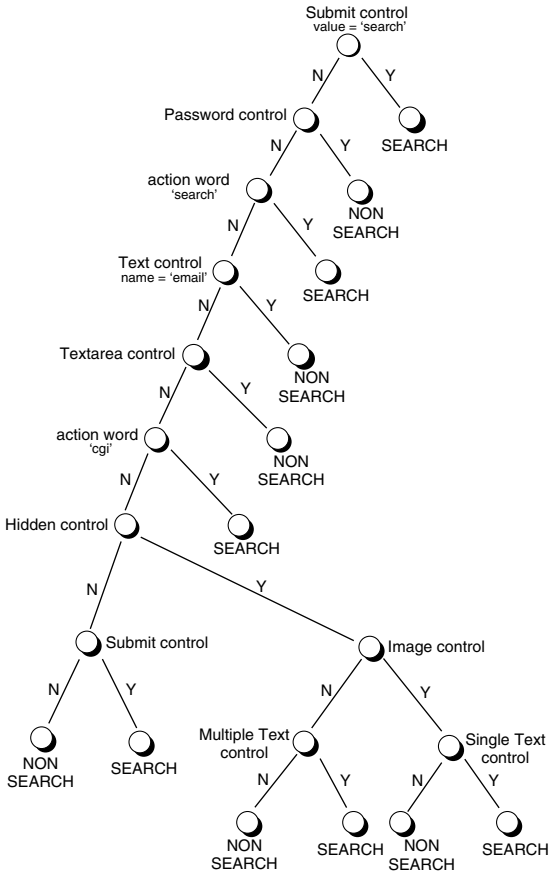


Fig. 5.2. Decision tree to discover search interfaces on the Web

To be able to automatically fill in forms, it is necessary to understand fields. Lage et al. (2002) make the assumption that labels are usually placed at the left on side or above to a form field. Zang et al. (Zhang et al. 2004) hypothesizes that there are regularities or (design patterns) among Web query forms and that they form together a visual language that can be parsed. They developed a tool converting an HTML query form into a set of tokens, each representing an atomic visual element of the form in a two dimensions layout. They propose to capture patterns such as proximity, adjacency, semantic relation between terms, etc. With a parsing algorithm, they achieved 0.80 precision and recall for the detection of forms.

When this is done, it is necessary to submit terms automatically via forms in order to generate responses. They are several other difficulties to deal with when trying to automatically fill forms (cf. Liddle et al. 2002).

Input to forms can range from text fields, radio buttons, check boxes, lists, etc. to any type of MIME encoded attachments. A logical request can be divided into several forms with state information captured on the server (cookies, hidden fields, values encoded into the base URL). Some forms rely on scripts to transform fields before submitting the form (range checking, other field validation and automatic calculation of certain fields).

If the vocabulary of submitted terms is well defined enough there is a reasonable chance to fetch most of the content. It is particularly the case when the domain of this vocabulary is limited or finite (like zip codes, dates etc.). In fact, one such a field suffices. For instance, a documentary gateway to French philosophers texts that would provide a date entry could be queried with all dates from 1100 to 2005 without any supposition on author's names nor titles of their writings. In 905 queries, one could be reasonably sure to get all the texts.

A limit of this approach is reached when query domains for all fields are too open or undefined to be systematically explored. In these cases, it is possible to use an other approach, which exploits result from a first query to extract new query terms, that will then be submitted and this iteratively. These methods presented in Callan and Connell (2001) is called query-based sampling and has been applied with some success (Agichtein et al. 2003) and (Barbosa and Freire 2004). Ntoulas et al. (2005) have proposed an adaptative algorithm to select the best keywords namely keywords that will return the most document. They show that with only 83 queries, it is possible to download almost 80% of the 14 millions documents stored in PubMed.

5.5 Crawler-Server Collaboration

5.5.1 Exposing Documentary Gateways to Crawlers

When collaboration is possible with the website producer, it is possible to expose hidden website content to crawlers in order to archive them. Here as often, archiving technology follows methods developed for search engines. They consist in generating a full list of documents or providing access to a service that can be automatically queried by crawlers. They range from generation of page, often hidden for human users, pointing to all objects of the website to dedicated querying protocols. Several such protocols have been proposed since the late nineties with limited success, except for some of them like the OAI Protocol. We will review the most significant of them below.

5.5.1.1 Hidden Links Pages

The simplest way to enable crawlers to get hidden content is to generate a list of links pointing to all documents of the hidden website. This applies particularly well for documentary gateways and can be completely hidden for normal users by hiding these pages from normal navigation, for instance by embedding hidden links to them in the homepage.

This has been made for instance by the Bibliothèque nationale de France and the National Library of Australia to expose their collections to search engines:

> A separate list of URLs was created for each of the Library's digital collections, namely pictures, maps, sheet music, manuscripts, and books and serials. Each collection includes thousands of items which are listed on a series of Web pages, each containing 100 links that resolve to the collection items. These pages use the robot directive "noindex, follow" to direct search engine harvesters to follow links to content but not index the list pages. The URL lists themselves are also dynamically generated with new content automatically added to the list as new items are digitized and made available on the Internet. (Boston 2005).

To be effective, a stable linking scheme has to be in place and all documents have to be linked. Brandman et al. (2000) calculated that such a mechanism could save up to 80% crawlers-servers exchange bandwidth. Several proposals to standardize this type of mechanism in a more formal way have been made since the 1990s, we will present one of them currently in use: the open standard RSS.[3]

RSS stands for RDF Site Summary (alternatively "Rich Site Summary" or "Really Simple Syndication." It is an extension of RDF that was originally conceived for the "MyNetscape" service. It is widely used today by blogs or news sites to provide a short list of the latest news and/or sites updates. Such a standard can be used to generate periodically an RDF file containing the URL and last modification date of all pages of the site. Crawler can then checked this file first and use it to crawl the site or compare it with their existing list of pages if the site has already been crawled and fetch only modified pages. Castillo (2004, p. 109 sq) presents such an implementation.

[3] See also Google Site Map: http://www.google.com/webmasters/sitemaps/ docs/en/ about.html (last visited May 2006).

Such a mechanism can be applied to list and expose any type of pages, be they linked from the rest of the website (usually the case) or not (hidden Web pages).

5.5.1.2 Protocol Level

A step further can be achieved with dedicated communication protocol like the Open Archive Initiative Metadata Harvesting Protocol (OAI-MHP), that has been proposed (Lagoze and Van de Sompel 2001) to expose metadata over HTTP using an XML syntax. Implementing at the protocol level permits a real communication i.e., with request and responses from the server, hereby multiplying the possibilities. It is for instance possible to query document by range of dates, type of documents etc. Crawler can directly communicate with the OAI server to get the list of documents with associated metadata (records). If the OAI compliant repository provides unrestricted access to documents, it is then possible to fetch them and stored them with their metadata in the archive. Some intermediary gateway services have also been implemented for crawlers that are not able to use OAI protocol to generate pages of links for all documents from an OAI server (Liu et al. 2002).

Even if this type of collaboration mechanism is in place, it should be considered in most cases as a complementary method of gathering content for archiving agency, as sites producers usually use them to reduce load on their server by targeting search engine crawlers visits towards recently updated pages. Whereas indexing crawls can offer to remain on the surface of sites and concentrate on freshness of their results, archiving crawls require integrity and completeness.

5.6 Archiving Documentary Gateways

In some cases the previous methods cannot be applied as they imply loosing the richness and structure of metadata associated with documents that cannot be mapped into a simple flat link. Imagine for instance a scientific image collection: archiving all the images without their metadata can be useless. In these cases metadata associated with images are as important as images themselves.

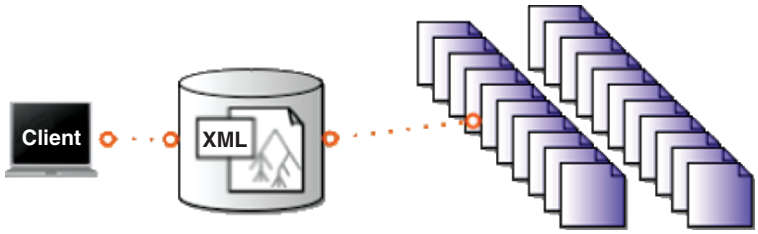


Fig. 5.3. Archived version of a documentary gateway. The database has been replaced by a XML file with all the metadata about documents. Mapping of links to the documents has been made

An alternative option to avoid implementation of a new protocol by the server is to directly extract metadata from their database and archive together with the documents in an open format (see Fig. 5.3). Bibliothèque nationale De France (BNF) has successfully applied this method to several hidden website in 2002. Though it requires collaboration from the producer, it is less demanding as it only requires an extraction, compared to

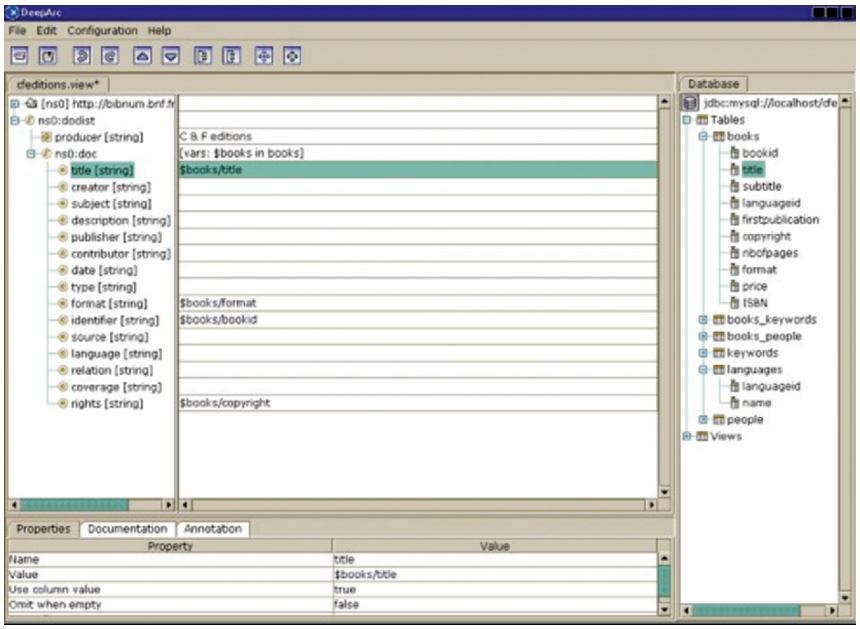


Fig. 5.4. DeepArc main screen with on the left column, the target schema model and on the right column, the database schema. One can graphically link the DB fields to the schema elements to prepare the extraction

implementing a new service. The main difficulty comes from heterogeneity of database systems, database scheme and mapping scheme to objects.

For alleviating the first two, Bibliothèque nationale de France has developed an open source database graphical extractor that can run on several database systems.[4] The tool can then export the database content into an XML document conformant to the schemas chosen by the archive. The producer who knows best his or her internal database schema can graphically map it with the target schema provided by the archive (see Fig. 5.4).

For instance, if the producer has used "AUT" as a field name in his database for author, he or she can graphically map the two with the help of a wizard. The wizard also enables the producer to filter out certain type information that need not be archived for privacy reason for instance. All the procedure can be made in a limited time and generate as an output a structured XML version of the database that can be exported to and preserved by the archive together with the documents described by this metadata.

The main difficulty of this approach comes from the naming and linking scheme of objects and how it can be translated in the archive environment. Identifiers of objects within the database can be arbitrarily complex and difficult to link to in the archive environment. If a directory structure has been used to organize objects, or if a script in the original environment generates the path to objects, the archive will have to construct its own linking scheme compliant with its naming scheme and structure. As the original linking mechanism can be arbitrarily defined, there is no general method to be applied here. A case-by-case approach is necessary to build a working archive.

This archive has to comprise its own HTML form querying the XML archived metadata and linking to the collections of objects.

Note that metadata can be injected in a traditional relational database management system to be queried easily in the archive. The important thing here is to preserve an XML version of the metadata originally contained in a database in order to make sure they will be readable in the future.

5.7 Conclusion

As we have seen, archiving the hidden Web is more difficult than archiving the surface Web. None of the methods presented can be considered as mature at the time of writing albeit several have been tested successfully. It will require further developments to be able to preserve in simple ways the

[4] DeepArc, an opensource database extractor. http://bibnum.bnf.fr/downloads/deeparc/ (last visited May 2006).

hidden Web. This possibility also depends on the technical evolution of the Web. But there at least two reasons not to neglect this portion of the Web. The first is that it is large and contains rich material that many heritage institutions are interested in. The second is the possibility that the Web's evolves towards architectures of information that resists traditional crawling techniques. The main counter-balance against this is the pressure that search engine puts on sites: to be indexed, they have to be crawled. But this might change if direct agreement and bilateral updating mechanism emerge, a good reason to continue to work on this issue for a preservation perspective.

References

Adams, K. C. (2001). The Web as Database: New Extraction Technologies and Content Management. *Online*, *March*

Agichtein, E., Ipeirotis, P. G., & Gravano, L. (2003). *Modeling Query-Based Access to Text Databases*

Barbosa, L. & Freire, J. (2004). *Siphoning Hidden-Web Data through Keyword-Based Interfaces*. Paper presented at the SBBD

Bergman, M. I. K. (2001). The Deep Web: Surfacing Hidden Value. *The Journal of Electronic Publishing*, *7*(1)

Boston, T. (2005). *Exposing the deep web to increase access to library collections*. Paper presented at the AusWeb05. The Twelfth Australasian World Wide Web Conference, Queensland, Australia

Boufkhad, Y. & Viennot, L. (2003). The Observable Web. *RR*

Boyko, A. (2004). Test Bed Taxonomy. *IIPC Reports,* 16

Brandman, O., Cho, J., Garcia-Molina, H., & Shivakumar, N. (2000). Crawler-Friendly Web Servers. *SIGMETRICS Performance Evaluation Review*, *28*(2), 9–14

Callan, J. & Connell, M. (2001). Query-based sampling of text databases. *ACM Transactions on Information Systems 19*(2), 97–130

Castillo, C. (2004). *Effective Web Crawling.* University of Chile

Chang, K. C.-C., He, B., Li, C., Patel, M., & Zhang, Z. (2004). Structured databases on the web: observations and implications. *SIGMOD Records*, *33*(3), 61–70

Cope, J., Craswell, N., & Hawking, D. (2003). *Automated discovery of search interfaces on the web*. Paper presented at the Proceedings of the Fourteenth Australasian Database Conference on Database Technologies 2003

Florescu, D., Levy, A., & Mendelzon, A. (1998). Database techniques for the World-Wide Web: A survey. *SIGMOD Records*, *27*, 59–74

Frankewitsch, T. & Prokosch, U. (2001). Navigation in medical Internet image databases. *Medical Informatics and the Internet in Medicine*, *26*(1), 1–15

Gravano, L., Ipeirotis, P. G., & Sahami, M. (2003). QProber: A System for Automatic Classification of Hidden-Web Databases. *ACM Transactions on Information Systems*, *21*(1)

He, H., Meng, W., Yu, C., & Wu, Z. (2005). *WISE-Integrator: a system for extracting and integrating complex web search interfaces of the deep web*. Trondheim, Norway

Hearst, M. (1998). Information Integration. *IEEE Intelligent Systems*, *13*(5), 12–24

HTTrack. http://www.httrack.com/

Lage, J. P., Silva, A. S. D., Golgher, P. B., & Laender, A. H. F. (2002). *Collecting hidden Web pages for data extraction*. Paper presented at the Proceedings of the fourth international workshop on Web information and data management

Lagoze, C. & Van de Sompel, H. (2001). *The open archives initiative: building a low-barrier interoperability framework*. Roanoke, Virginia, United States

Lawrence, S. & Giles, C. L. (1999). Accessibility of Information on the Web. *Nature*, *400*, 107–109

Liddle, W. S., Yau, S. H., & Embley, D. W. (2002). *On the Automatic Extraction of Data from the Hidden Web.* Springer, Berlin Heidelberg New York

Liu, X., Maly, K., Zubair, M., & Nelson, M. (2002). *DP9 - an OAI gateway service for Web crawlers*. Paper presented at the Second ACM/IEEE Joint Conference on Digital Libraries

Ludäscher, B. & Gupta, A. (1999). *Modeling Interactive Web Sources for Information Mediation*. Paper presented at the Intl. Workshop on the World-Wide Web and Conceptual Modeling (WWWCM'99), Paris

Marill, J., Boyko, A., & Ashenfelder, M. (2004). Web Harvesting Survey, 10

Masanès, J. (2002). *Archiving the deep web*. Paper presented at the 2nd International Workshop on Web Archives (IWAW'02), Roma, Italy

Mohr, G., Kimpton, M., Stack, M., & Ranitovic, I. (2004). *Introduction to Heritrix, an archival quality web crawler*. Paper presented at the 4th International Web Archiving Workshop (IWAW'04), Bath, UK

Ntoulas, A., Zerfos, P., & Cho, J. (2005). *Downloading textual hidden web content through keyword queries*. Denver, CO, USA

Raghavan, S. & Garcia-Molina, H. (2001). *Crawling the Hidden Web*. Paper presented at the Proceedings of the 27th International Conference on Very Large Data Bases

Roche, X. (2006). Copying web sites. In J. Masanès (Ed.), *Web Archiving.* Springer, Berlin Heidelberg New York

Storey, M.-A. & Jahnke, J. H. (1999). *Web site evolution – Towards a flexible integration of data and its representation*. Paper presented at the 1st International Workshop on Web Site Evolution (WSE'99), Atlanta, USA

Zhang, Z., He, B., & Chang, K. C.-C. (2004). *Understanding Web query interfaces: Best-effort parsing with hidden syntax*. Paper presented at the Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data

6 Access and Finding Aids

Thorsteinn Hallgrimsson

National and University Library of Iceland
thh@bok.hi.is

6.1 Introduction

In libraries preservation and access are symbiotic in the sense that unless the published material, traditionally in printed form, is collected and preserved for the future access becomes unimportant, and conversely if access to the preserved material is not provided the preservation effort will ultimately become irrelevant. Therefore user access to a Web archive is of paramount importance not only for making the best use of the archive but also for those responsible for it to monitor and control whether harvesting of the Web is returning the desired output. It is however neither obvious what kind of access is needed nor who will be allowed to have access, because the web archive represents an independent digital medium and it will be difficult to provide the same kind of access as to traditional library collections.

Library collections have until very recently been mainly physical like books, journals, and manuscripts with later additions of analogue media like records of music and video collections. Since about 1985 collections of digital material on physical media like computer discs and tapes have been added to the collections. Access to those collections has been provided by traditional bibliographic cataloguing methods and by patrons either borrowing the physical media or accessing the digital collections through the Internet.

Access to archived Web material poses different problems and demands solutions that are based upon the same methods and procedures that are used in the public Web and the Internet. Therefore, organizing and providing effective access to the collected content requires an adaptation of existing methods and tools, and particularly to take into account the time dimension of Web archives with the several versions of a Web document.

In addition to the need for accessing the archive in a similar way as the public Web, there are other issues, both functional and technical that belong to the category of access. The archive will contain a vast amount of data and various statistics and information can be extracted from it. The national libraries could either alone or in collaboration with others define a new role for themselves by providing new services, perhaps similar to what some of the national statistic institutions already do, based upon the data mining techniques discussed in Chap. 7 of this book, "Mining Web Collections."

Potential users of a Web archive will have very different interests and expectations as to what data and information, access to the archive should provide. Just as for traditional material the content of Web documents will be of interest to the majority of users, are they the general public, journalists, scholars, or researchers. For the scholarly community it will be important to preserve Web documents that have been cited or referred to. If they disappear the citations and references become meaningless and the original paper loses some of its value. For those who study the Internet as a medium and want to follow its development the archive will be invaluable.

Another aspect that must be kept in mind is that in the long term, 20–50 years from now, users needs and subjects of interest will be different from what they are now. Then the archive will be a resource on par with today's research resources like journals, letters, books, and media like radio, television, and movies. Although one can make reasonable assumptions that certain topics and subject will be important in the future it is impossible to predict all of them. For the short term this is easier and the IIPC[1] has identified many areas that could be of interest, like providing an URL for a citation that refers directly to a persistent copy of the original site, documentation of the first occurrence of a certain document on a website and how it has changed, studying the evolution of certain websites or in an extreme case try to reconstruct a lost website.

From the point of view of collecting and preserving the WWW our understanding of the medium and its contents leaves much to be desired. Currently access (navigation, indexing, searching) to the documents in the archives is secured for research purposes, but general access to the documents depends on the copyright and legal deposit legislation in each country and is often very limited. This must change in the future. The archives contain multimedia documents and currently it is only possible to index text. For other types of documents like images and sound files, indexing is limited to harvester generated metadata plus textual information in the file header.

[1] International Internet Preservation Consortium (http://netpreserve.org).

As already discussed, there is a need for providing different methods for accessing a web archive. They can be a simple URL-based access, a content-based access via a Search Engine like Google and Yahoo, or Data Mining and analytical tools for extracting information. In all cases a suitable user interface must be provided. Every current access tool needs to be improved and new tools must be developed. Hopefully increased awareness and activity in Web Archiving coupled with advances in technology and international cooperation will in the near future enable the national libraries of the world to both preserve the Web and to provide the needed access to the Web archives.

6.2 Registration

Libraries, archives, and other collectors depend upon registration of their material for finding specific items in the collections. This has evolved from ledgers, lists, and cards to digital records and digital repositories. In all cases an identifier is needed and often many identifiers are used. Libraries and archives have developed very elaborate rules for cataloguing and registration of their material and the result is usually referred to as bibliographic data. For digital documents this registration is called metadata. It is indisputable that good metadata can greatly enhance access to digital collections and it is, therefore, natural to evaluate whether it is possible to use bibliographic methods to create suitable metadata for a Web archive. The harvesting software usually adds fields of technical metadata but otherwize it is quite rare that authors of Web documents have described their resources with metadata that would help in finding the documents.

The enormous quantity and variety of documents and formats that are present in the Web makes it very difficult if not impossible to use traditional methods for cataloguing/registering the information in such a manner as to make it accessible. In the PANDORA project, administered by The National Library of Australia, selected websites are collected and stored in the Australian Web Archive. Each title collected (a title is defined as a collected entity that is either a website or part of a website) is catalogued and incorporated into the Australian National Bibliography. At the end of 2005 the PANDORA archive contained about 10,000 titles and 28 million files/documents and its total size was about 1 TB.

In the Minerva project the Library of Congress collected Web material pertaining to the elections 2000, September 11, 2001, the Winter Olympics 2002, and others. Each collection has a limited scope and it was intended

to create catalog records (MARC) for each of the websites. For the elections and the Winter Olympics this was possible while the number of websites is about 5,000 with about 90 million Web documents, but for the September 11th and the 9–11 Remembrance, the number of websites is almost 32,000 with about 330 million Web pages and for those collections only about 2,300 were selected for cataloguing.

The experience from the Pandora and the Minerva projects shows that except for a very small Web archive traditional cataloguing is not an option. Therefore, it must be evaluated whether it is possible to make use of descriptive metadata that is imbedded in the META tag in HTML Web documents to create registration records that can be used for access. In theory it is possible to use it to create a subject list indicating the content of a Web page, but at present a very low percentage of Web documents contain valuable descriptive metadata. In addition the structure and contents of the descriptive metadata that is provided cannot totally be relied upon because of the abuse of the META tag by webmasters who try to artificially raise the relevancy of a page by loading it with terms unrelated to the actual content of the page. In response a lot of work has been done to standardize the META tag content in projects like in the *Dublin Core Metadata Initiative* and the *Digital Object Identifier* (DOI) system for identifying content objects in the digital environment. Those initiatives will certainly to some extent improve the reliability of descriptive metadata and it will probably be most apparent in a somewhat controlled environment like the scholarly community and within universities and research institutions. Although those documents will be important the volume as a percentage of Web documents will be very low and, therefore, probably not be useful in providing general access to an archive. Still the descriptive metadata can be very useful in providing access to specific collections, like within a university, and it can be used both to create traditional catalog records pointing to the archive or in indexing providing very exact references. Digital documents can in addition to descriptive metadata have embedded various kind of metadata that may be of use in accessing the archive but is to diverse to be of consideration at this stage in the development of Web archives.

During the harvesting of Web documents it is important to collect as much information as possible about the origin of the documents and when they are harvested and ingested into the archive. This can be called provenance metadata and is needed both for ensuring preservation and authenticity of the Web documents and for access to the archive. The IIPC has defined metadata types to be created during web harvesting and has where possible complied with the Open Archive Information Standard (OAIS). The metadata set covers information about the document itself

(URI, size, format, timestamps, checksum, etc.), harvesting activity (harvester name and address, server address harvested, etc.), transaction (harvester request, server response, etc.), selection process, and change history. A Web archive will as time goes by consist of documents harvested from selected websites and very often the exact same document will be harvested multiple times. Some documents will have slight changes, some will disappear and new ones will be added. In order to keep track of individual documents and how they change every document (file) will be given a unique identifier including a timestamp. Ideally the collected provenance metadata should specify what transformations of the document have been made, but this cannot be relied upon.

6.3 Indexing and Search Engines

From the above it is concluded that registration or cataloguing is not a practical solution for providing access to a Web Archive. This is very much in accordance with common practice for the public Web where most users find Web contents and gain access to the websites where they are available, by using one of the available search tools like Google, Yahoo, and others. This also corresponds with the decisions made by those that have collected and stored a large number of websites and Web documents like the Internet Archive and the Nordic countries, and with one of the conclusions of the final report for the Minerva Project that states: "The Library should rely on automatic indexing as the primary means for information discovery of websites and content within websites. The full text of all textual materials should be indexed on a periodic basis and searchable by users," and "The Library should not invest in extending MARC cataloguing or other manual methods of cataloguing to websites, except for sites of particular importance to the Library and its users."

An index based on a Web archive will be cumulative; that is, depending on the harvesting policy it will contain a reference to every document that has been harvested and stored. In records built by the indexing software there will be URL links to the archived documents.

Most of the present Web access tools are based on automatic indexing, where either a part or the whole text of documents harvested from the public Web is indexed. There are various types of automatic indexing and when the Web was gaining in popularity three different kinds were used: a simple back-of-the-book style index like in many scholarly books where keywords are linked to a page or chapter (for the Web keywords are linked

to a URL), indexing coupled with subject trees of reviewed sites, and indexing plus search engines (see The American Society of Indexers).

Simple back-of-the-book indexing does not work very well for a Web archive mainly because of the enormous volume of documents to be indexed leading to a great number of results or hits for most keywords. This in turn makes it very difficult to determine the relevancy of items (URL's) found via the search and to order (rank) them meaningfully.

Some Web search tools like Yahoo and Magellan combine automatic indexing with manual labor to decide which categories and keywords fit the site, providing a kind of reviewed subject based site index. This can give good results for the selected categories of websites.

The third indexing method combines automatic indexing of the full text with advanced computer applications called Search Engines for searching in the index. Google is a prime example of using this method for the public Web, but it combines complex indexing and sophisticated Search Engines to provide very reasonable search results.

Currently indexing of a Web archive is mostly based on what text is available for extraction in a Web document. For some resources – images, for instance – indexing will be limited to harvester generated metadata plus textual information in the file header. However, a lot of research is being done with the purpose to use computer programming to index sound, images and video/movies and when this becomes practical these parts of the web archives will be indexed.

6.3.1 Relevancy

The term "Relevancy" means that results are sorted, or ranked, according to certain criteria and algorithms to determine the relevancy of items found via the search. Criteria can include the number of terms matched, proximity of terms, location of terms within the document, frequency of terms (both within the document and within the entire database, document length, and other factors.

Of course, no Search Engine measures these components in the same manner. For example, all Search Engines remove common words from a page, called stop words or filter words. Those are common words (a, an, of, or, and the) that the search engines ignore during a search because they slow down the retrieval of search results without improving accuracy.

They also rank results in different manner. Google has been the first to use the inlinking to pages to rank results.

PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

6.3.2 Integrity

For the public Web it can be difficult for the user to be aware whether the indexing and Search Engine methods have been tampered with in order to place certain websites higher up in the ranking than they actually deserve. This can be the result of an inappropriate bias in the product but it can also be because the methods used do not cope well with those who use dubious methods to get a higher ranking for their websites. There is a constant battle between those trying for a higher ranking and the developers who try to achieve the integrity needed for providing the user with an honest and objective way to find high-quality websites with relevant information.

6.4 Access Tools and User Interface

An access tool in this context is a tool that enables the user to access a Web archive and retrieve archived objects. A user interface is the users window into the archive with facilities for defining what should be retrieved and for rendering the results. It is quite conceivable that an access tool must support a variety of requirements and could, therefore, have different user interfaces.

As discussed in the introduction to this chapter different methods and access tools are needed for accessing a Web archive depending on who the users are. In order to better understand their needs it makes sense to define certain user groups and the functionality required for each group. Each archive will probably have its own special regulations for access and what data can be retrieved. This can range from unrestricted access to a very restricted one, where a few number of authorized users will have access, and each archive must design its own access regulations. It is very important in this context to recognize that even if access is restricted there is still need for basic access tools and one could argue that if the archive is to be of use for a user group like researchers, however few they may be, it

is important to define a full set of the functional specifications for access to, and retrieval from the archive.

Another important user group is the administrator of the archive. The administrator must be able to verify that the harvested documents in the archive are in accordance with the adopted collection policy. This pertains both to coverage (subjects, sites, categories of documents), and exhaustiveness (what percentage is harvested). Therefore, in the following discussion the relative importance of access tools will not be evaluated.

Before discussing access tools it must be recognized that archives of material published on the Internet are still in their infancy; the oldest archive, that of the Internet Archive, is a mere eight-years-old. The Access Working Group of the IIPC has produced a Use Case document discussing the different user requirements for access and in the introduction it is stated that:

> It is conceivable that just as the advent of the Internet has forever changed the way people publish and exchange information, the availability of archives of Internet material will result in new and innovative usages of archives. Much of the current effort related to Internet archives focuses on collecting and preserving the information. The success of the Internet however is based on the easy access to information. It seems reasonable to assume that the ultimate success of any Internet archive will be measured by the means of access the archive provides to the preserved material.

A Web archive reflects the public Web and is a collection of objects that have been copied from the Internet by harvesting software. The majority of these objects are html and text files, but images in various formats are abundant. In addition there are sound objects and moving images (video and movies) and many other types like programming scripts. An archived Web document usually consists of several objects usually a HTML file (one object) and often inline images (other objects within the same document).

Documents retrieved from a Web archive they should be rendered in the same manner as when they were harvested in the public Web. This means that text should be in original format and that music should sounds like the original. There are several inherent difficulties in achieving this because the Web is becoming more complex and often there are interactive components, programming scripts like Java scripts or special rendering programs, so-called plug-ins, that are required for rendering the document. All of those must be harvested at the same time as the document or made available either in the archive or in the user interface.

The public Web only has one version of a given website, whereas the Web archive will be cumulative, i.e., depending on the harvesting policy it will include several versions of a website and contain all documents from

that site since the harvesting began. This introduces the element of time and the issue of duplicate documents. The time element is one of the major advantages of a Web archive because it makes it possible to analyze how the Web changes over the years and it must be recognized in the access tools and in the user interfaces that apply (see Chap. 7 on "Mining Web collections").

Duplicate versions of the same document are a challenge because it can be very tedious and confusing for a user if he is presented with many identical documents during access and various efforts are being made to develop methods to determine whether a document has changed or not.

In order to access a Web archive both an access module and a suitable user interface must be provided. They are closely interrelated but it is very conceivable to define and implement multiple user interfaces that use the same access module. In general it can be stated that an access module should enable the user to search and retrieve documents in the archive based on the functional requirements specified by the user(s), libraries, or others on a standard interface between the access module and the archive.

It has already been mentioned that the potential users of a Web archive are just as diverse as the users of the public Web and because this is an archive covering a certain time span they will have additional requirements. In order to illustrate this the IIPC, in has identified several scenarios for using the Web archive, i.e., what kind of access and services will be provided. They by no means cover all aspects, but prove the need for different tools like a simple URL-based access, a content based access via Search Engine like Google and Yahoo, or data mining tools and analytical tools for extracting information. In all cases a suitable user interface must be provided. Some are useful to all users, others are more specialized.

Users can be grouped in many ways but it is important to make a distinction between external users who come from outside the institution that administers the archive and internal users coming from within the Web archiving institution. In addition there will be users who want to extract information from the archive by processing it as a whole or parts of it.

For external users the use of a Web archive differs from use of the public Web. The user may be looking for documents that are no longer available in the public Web, want to explore the development of certain documents or events or need information about what was happening at a certain time. For a casual user the list is endless but some examples can be: a journalist looking for documentation of a news item, a person wanting to examine how certain regulations have evolved, a person looking for the

first appearance of a topic, a marketer studying a certain product, or just interest for the past. For a researcher or a scholar the interest may be certain subjects or themes and how they evolve, the medium itself or parts of it. Of vital interest to the academic community is to ensure that citations that refer to Web documents will remain valid. This implies both a persistent identifier for the reference and a secure archival of the original document. In an extreme case a user may want to reconstruct a lost website.

Internal users are institutions that create and maintain an institutional, national, or international Web archive. This is a difficult and demanding task and a successful result will reflect the variety and complexity of the public Web. The Web Archiving institution must be prepared to provide information about the harvesting activities and the archive not only because of the freedom of information acts that many countries have in their legislature, but also for common courtesy and sense. The institution must understand the harvesting activities, both on a total scale and also for individual websites that are harvested, and be prepared to provide information about this to the owners of individual domains. It must also be able to provide information about the composition of the archive.

Those who design and provide access to the archive and who control the harvesting activities, will need data and information to understand its composition and will need access tools to help with that.

General statistics are needed and this requires a significant degree of analysis of document content both in total for the archive and also over a period of time. It is not possible to totally rely on the data and information that can be collected and stored during harvesting, both because of the iterative and cumulative process involved, and because multiple harvesters may be used. Therefore statistical data and information must be extracted from the archive and for that tools must be developed.

In this context it is important to understand the frequency of updates and the addition of new material, or completely new content. This data will eventually help to avoid spending resources unwisely, and serve as a checklist to ensure that important information is not overlooked. This will ultimately depend on decisions about harvesting policy but once that is determined future resource requirements must be estimated, e.g., for archive space, communication capacity, processing capacity, frequency of updates and the addition of new material, or completely new contents.

Information about "document quality" is needed to provide information for post processing of documents and to ensure that as much as possible of the archive can be correctly interpreted and rendered in 100 years. It must be known what formats, character sets, document type definitions, etc. are

in common use in order to make sure that future software will be able to handle as many of those as possible.

Users who want to understand the nature and development of the Internet need statistics and analytical and data mining tools that can be applied for various purposes to process the archive or parts of it. Many will serve multiple purposes This subject is discussed in Chap. 7 “Mining web collections.”

To fulfill the many requirements for accessing and working with a Web archive many different solutions are needed. A simple access based on a URL will in many cases be sufficient, for other purposes a sophisticated full text search and retrieval is required. Many users will need to process either the entire Web archive or certain sections of it. Every current access tool needs to be improved and new tools must be developed. Hopefully increased awareness and activity in Web archiving coupled with advances in technology and international cooperation will in the near future enable to both collect and preserve the Internet and to provide the needed access mechanism to the Web archives.

6.4.1 Access Rights

At the beginning of this chapter it was stated that access to a Web archive is of paramount importance not only for making the best use of the archive but also for those responsible for it to monitor and control whether the harvesting is returning the desired output. Therefore, it is assumed that the archiving institution will have unrestricted access to the archive. When considering access for other users the nature of the archive must be considered. The Web is a medium of its own and it is quite different from media like books, journals, television, radio, and others, but it contains material that could belong in all of those. It has both the character of a library and an archive and the collection must be available for research and should be available for other use as well. Whatever the use may be it must not violate the privacy of individuals or the legitimate interests of those who published the material on the public Web and it is not intended to replace use of the public Web.

It is not possible at this stage to predict what kind of access will be provided in the future and this will often depend on the legal foundation of the Web archive. This varies from country to country and in addition the laws of copyright and personal data may apply as well. Therefore, potential policies and regulations will not be elaborated except to state that provisions must be made for controlling the access to the documents

according to those laws. In the Minerva report this is explained as follows:

The difficulty is that there are no good parallels to use in setting such policies. An initial view might be that, since these materials were made available on the web with open access, the copyright owners expect them to be read and studied without restriction. For many sites, probably most, this is a valid assumption. The people who mounted them hoped that they would be read and will be pleased that they are available for future scholars. Other copyright owners, however, may be less enthusiastic. Materials may have been mounted for a specific purpose at a specific time; they may have included errors that were subsequently corrected. The potential for violations of privacy by data mining also needs to be taken seriously. While the aim should be to encourage widespread use of the collections, totally unmonitored access to all collections is probably inappropriate.

Conversely, a sad mistake would be to fall back on the habits of the past and require users to be physically present at the Library of Congress. This imposes burden on scholars by forcing them to travel unnecessarily and places burden on the Library to provide computing facilities for researchers to analyze the materials.

In addition to the above there are many commercial and "fair use" issues that must be considered when providing access to an archive and potential users will be of different categories. For most of the harvested material access in the public Web is unrestricted, but some of it is only accessible for a fee or restricted by some other means. It is obvious that it will never be accepted that harvested restricted data can be accessed everywhere by all. An example of this is a newspapers article index that requires payment for use. There are plenty of other similarities.

The users needs and requirements differ and this must be considered. Legitimate researchers should be provided with more access rights than the general public, and this could apply to the law enforcement as well. Currently there is very little practical experience for how access to a Web archive will be.

On one extreme it can be very restricted and access will only be allowed for research and statistical purposes in a "secure network." The services will be offered at computers that are placed at the domicile of the archive or at entrusted institutions like the national library, or university libraries, that are connected to the archive through secure communication. In Sweden the Web archive is accessible to everybody but only at the National Library premises.

On the other extreme, access can be through a public website like for the Internet Archive collection where access is open to everyone six months after the material is collected. (Still the Internet Archive removes access to Web documents in the collection if the owner asks for it). See Chap. 9 on the Internet Archive.

An in between solution would be to offer registered users a login access through the public Web. This is distinct from the fully public website because it would enable scholars and researchers to use the collections where it is most convenient for them to do research (See DACHS example in Chap. 10).

From the point of view of collecting and preserving the Internet our understanding of the medium and its contents leaves much to be desired. Currently access (navigation, indexing, searching) to the documents in the Web archives is secured for research purposes, but general access to the documents depends on the Copyright and Legal Deposit legislation in each country and is often very limited. This must change in the future.

Every current tool needs to be improved and new tools must be developed. Increased awareness and activity in Web archiving coupled with advances in technology and international cooperation will in the near future enable the national libraries of the world to collect and preserve the web like the printed collections of today.

Hopefully a practical consensus will be reached for access to the Web archives because they will become increasingly important as time passes and an ever growing amount of valuable material and data is only available there.

6.4.2 Technical

Although it is beyond the scope of this book to discuss all the technical issues involved in accessing a Web archive it is appropriate to mention a few of the issues, challenges, or problems that must be addressed. In this context it is very important to keep in mind that a Web archive can be very large and contain several billion records. This must always be considered when developing efficient Web archive applications that can cope with this number of records.

As Web archiving becomes more common the archives will proliferate. One of the missions of the IIPC is to foster the development and use of common tools, techniques, and standards that enable the creation of, and access, to international archives. The IIPC working groups have discussed many of the complex issues encountered in Web archiving and have suggested a system architecture for this and identified the interfaces between

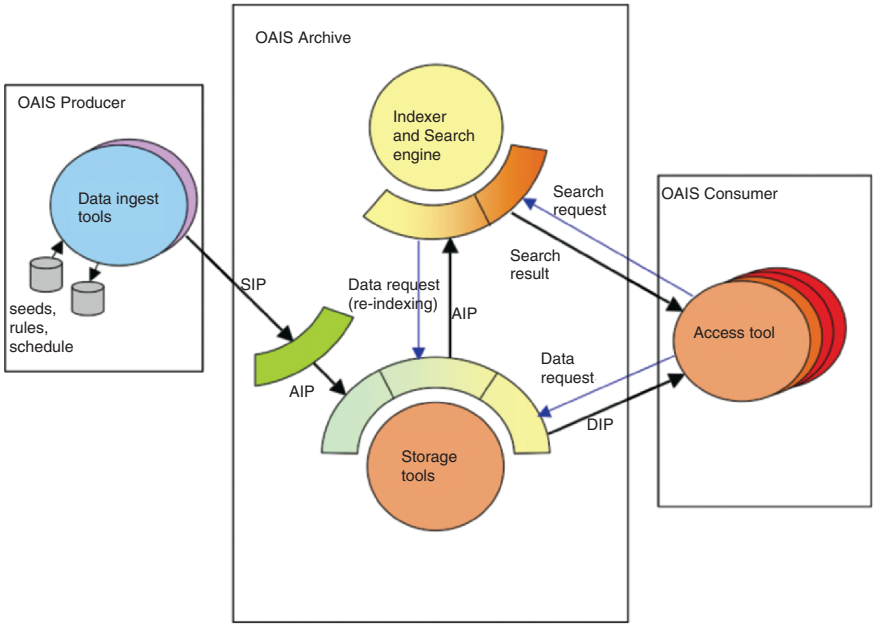


Fig. 6.1. The IIPC functional architecture based on the opean archive information system (OAIS) model

the main components. Web archives will differ in organization and structure and the access tools developed should ideally strive to be independent of an archive. This can be achieved by defining a standard interface between the access module or tools and the archives (see Fig. 6.1).

In its simplest terms accessing a Web archive involves searching an index and based upon the results to navigate within the archive. The search functionality usually returns a list of references that link to the archive and serves as the entry point for the user. The archive records have a header record with an imbedded unique key that most likely will be based on the URL plus at least a time stamp. It is relatively simple to index the header records and most of the present Web archives utilize this method for providing access to the associated documents. Providing an index of the full text in the archive in order to provide same functionality for access as most users of the real Web are used to, i.e., full text indexing combined with a Search Engine, has proven elusive. As mention before this is a very complex task and it requires a great deal of computing capacity. There are several commercial and/or proprietary applications available for indexing Web documents but they are not designed for a Web archive. The IIPC members at least have felt the need for development of indexing applications and now a very promising open-source Web search engine

called Nutch is being adapted under the name NutchWAX "Nutch + *Web Archive eXtensions*" so that it can be used to index and search Web archives. Currently it is capable of indexing about 140 million documents in a reasonable time but obviously it needs to be scaled to handle billions of documents. This development ensures that it will be possible to provide the same kind of access to Web archives as to the real Web.

Navigation within the archive is very important and the user should, e.g., be able to follow links, and move to later/earlier versions of a document by using various methods. Some are directly parallel to the public Web like surfing and following links. Others are archive related like enabling bidirectional or reversed linking, following "what documents refers to this page?" or finding "from where in the Web can I get here?." Implementing this is not trivial and one issue that must be dealt with is to ensure that when a user is following links in the archived documents the link is rewritten to point inside of the archive instead to the original location on the live Web. If not the user could suddenly be looking at documents in the live Web.

As mentioned earlier it is certain that as time passes there will exist in a Web archive multiple versions of a website and Web documents in the archive. Some will be identical (no change between harvesting) but others will differ and although ideally there should only be one copy of identical documents it is very difficult to define when a document does change. Identification and elimination of duplicates is of importance for reducing the size of an archive and it will become even more important in the access tools when there are a lot of duplicates in the archives. It will then be very tedious for the user to work through the duplicates.

A very common method for determining if a digital document or file has changed is to calculate a unique code for each document or file, e.g., by using the MD5 algorithm that takes as input a message of arbitrary length and produces as output a 128-bit "fingerprint" or "message digest" of the input, commonly called a "content digest." If two files differ by a single bit, they have different codes and if several files have the same code the probability that they are not identical is essentially zero. This method works very well for documents that do not have interchangeable parts like dynamic images or a layout that can change. Currently many Web documents especially HTTP documents have those characteristics and, therefore, additional methods must be used. There are fields in the HTTP header that can be relied upon correctly indicating that a change has occurred, but they are much less useful in accurately indicating no change when none has occurred. Furthermore HTML documents can be very volatile and accurately detecting change in HTML documents is very tricky as they are frequently generated on demand, and it is probably not

feasible to do this during harvesting. Sound and video formats are much more stable while they are typically not generated on demand and as a result most changes will require some user action.

Harvesting statistics show that text documents, mainly HTML, comprise between 70 and 75% of the number of documents and about 33% of the stored data, and nontext documents, such as images, movies, audio, and encoded text documents (PDF, MS Word, etc.) account for less than 25% of the documents gathered but more than 60% of the data stored. Therefore for reducing duplicates the easiest method is to find duplicate sound and video documents. A Heritrix module called the DeDuplicator has been developed and it has proven quite successful during harvesting in using both the content digest and the header information to decide if nontext documents have changed thus reducing both the amount of storage and the number of stored documents.

Since it is not feasible to detect change in HTML documents during harvesting this leaves the problem of working through many duplicate versions of HTML documents during access. A possible solution could be built into the user interface allowing the user to select documents from the list of results from a query and input them to an application for comparing HTML documents. The application would for each item on the list indicate the probability that the document has changed, allowing the user to act on that. The advantage is that user can decide if this is warranted and the amount of documents to be processed would be reasonably manageable.

In the public Web there are several hundred document formats although a few like HTML, GIF, JPEG, TIFF, and PDF are predominant, Most Internet browsers can render the common formats like GIF, JPEG, and TIFF but others either require special software packages, e.g., PDF and MSword, or so-called plug-ins. Most of the plug-ins are available free of charge and some are installed together with the browser.

Rendering all document formats is a challenge and will in the future become even more difficult. A long term solution has not been found but the user interface should provide information about the file format and what plug-in or software is needed to render it. Optionally there could be a link to a help file if the user's computer is not equipped to render the file.

6.5 Case Studies

There are not many public tools or applications for accessing Web archives. This is not surprising since there are only a handful of Web

archives available to the public. By far the largest with over 60 billion resources is the archive collected by the Internet Archive and it is therefore appropriate to look at the Wayback Machine, the tool used for accessing that archive. Currently the IA is working on a new open source access tool, the "New Wayback," designed to eventually be its standard tool for querying and rendering archived content. It will have an Opensearch API, and preliminary development has been completed to allow requests to be satisfied using NutchWAX full-text indexing and Search Engine. Another tool that will be looked at is the WERA (WEb ARchive Access) an application that gives the possibility to do full text search in Web archive collections as well as an Internet Archive Wayback Machine-like access, and easy navigation between different versions of a Web page. Both the "New Wayback" and the WERA are supported by the IIPC and they will most likely become very similar in features in the future.

6.5.1 The Wayback Machine

As already mentioned the Internet Archive in 1996 began archiving the Web. Soon after the original idea for the Wayback Machine was formed, but it was not made available to the public until the fall of 2001. It was developed for the Internet Archive by Alexa. The Wayback Machine has a very simple user interface for accessing a Web archive and rendering archived Web pages belonging to different versions of websites. This is illustrated in the following example.

The user starts by typing in the Web address of a site or page. The results is a list of all the dates when this site has been harvested (see Fig. 6.2).

The user can narrow the search to a particular year if necessary (see Fig. 6.3).

Some of the in-line images are missing but the links that are available on this page can be followed and they point to Web documents archived at as close a date as possible. Using this method the user can follow the evolution of the website of the National Library of Australia and any other website that has been harvested.

The Wayback Machine is an excellent tool for browsing in a Web archive but it does require the user to know beforehand the URL for the website to be found. This may not be a limitation because the URL can usually be defined. Keyword searching is not currently supported.

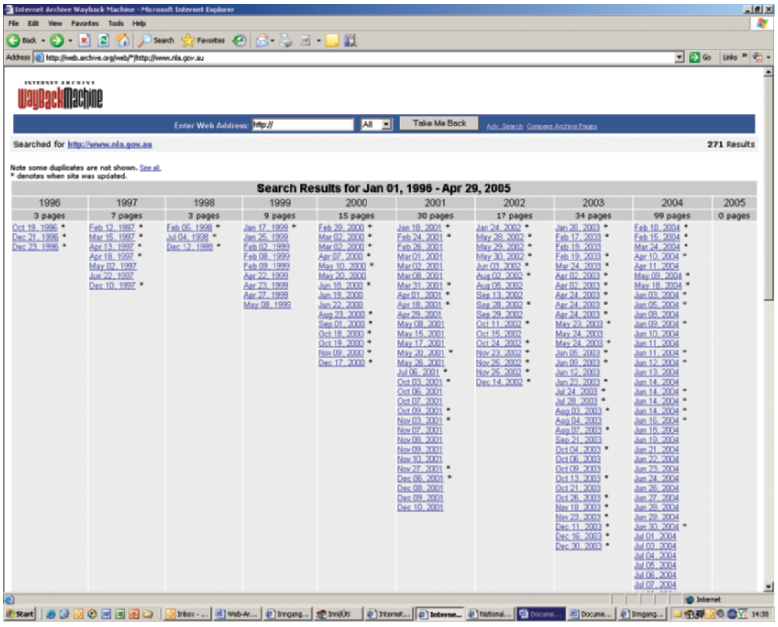


Fig. 6.2. Presentation of results for given URL in the Internet Archive's Wayback Machine

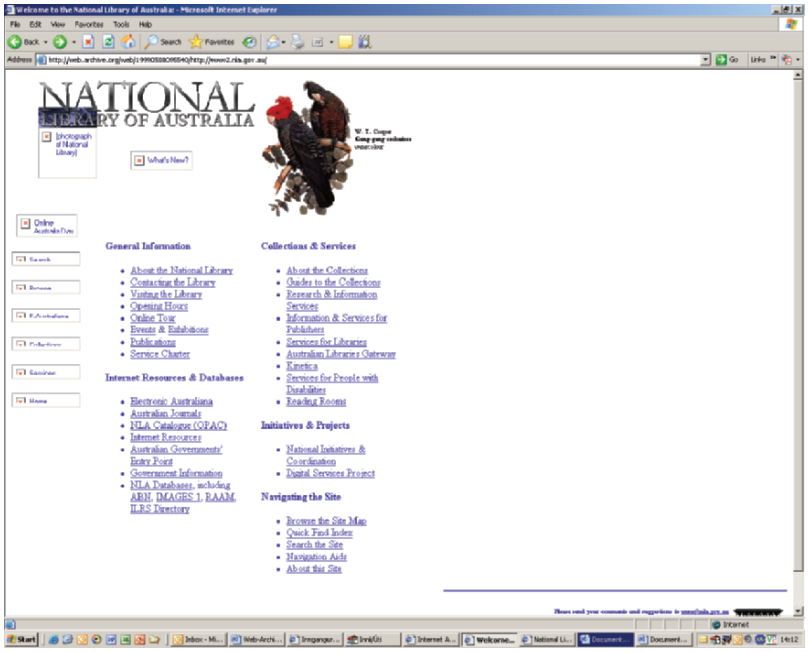


Fig. 6.3. View of the archived version of the NLA homepage, dated May 1999

6.5.2 WERA

WERA is an archive viewer application that gives an Internet Archive Wayback Machine-like access to Web archive collections as well as the possibility to do full text search and easy navigation between different versions of a Web page. WERA is based on and replaces the NWA Toolset. WERA needs the NutchWax Search Engine to offer full-text search and page navigation. The user interface allows the user to search, browse, and navigate in the archived Web documents and render the result in the same manner as using Internet Search Engines like Google in the public Web.

Following is an example showing the use of WERA for accessing a Web archive that contains Icelandic websites harvested weekly since March 16, 2006 in order to document the municipal elections that will be held on May 27, 2006. The search word is "Reykjavíkurborg," i.e., Reykjavik city (see Fig. 6.4).

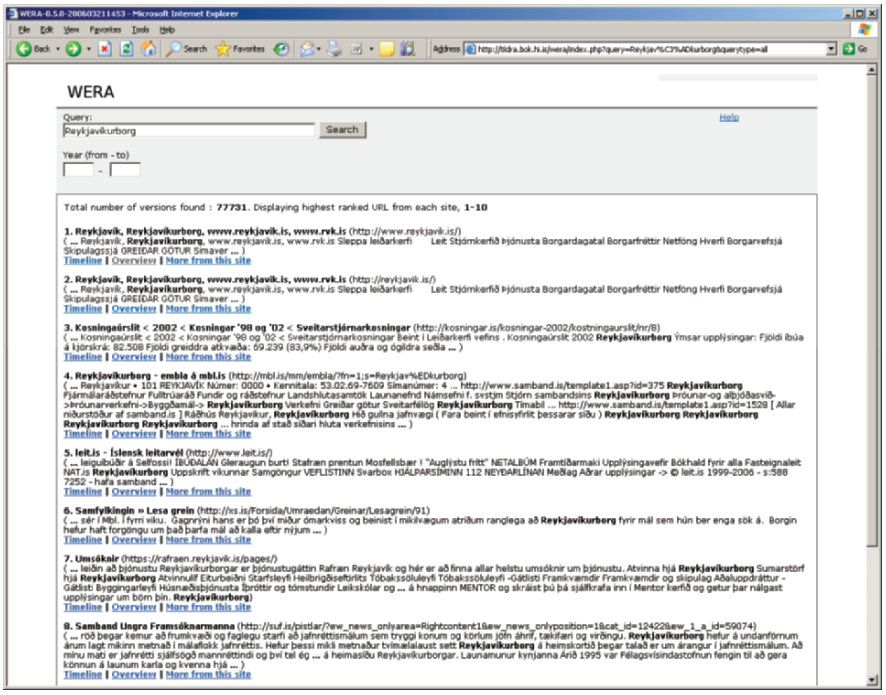


Fig. 6.4. Result page for a full text search page using WERA and NutchWAX

For each document the user can choose to view an overview of the dates of harvesting or to view a timeline. An overview for document "Reykjavík, Reykjavíkurborg, www.reykjavik.is, www.rvk.is (http://www.reykjavik.is/)" was chosen. The overview shows six different versions and from this the second one was chosen (see Fig. 6.5).

The English version of the document is rendered and at the top there is a timeline with four clusters showing when this document was harvested with an arrow pointing to the rendered one. The user can navigate forward and backward for this document using the timeline. In the example shown the time scale is months, but it is possible to change the time scale to show years, days, hours, or minutes. The embedded links can be followed as well.

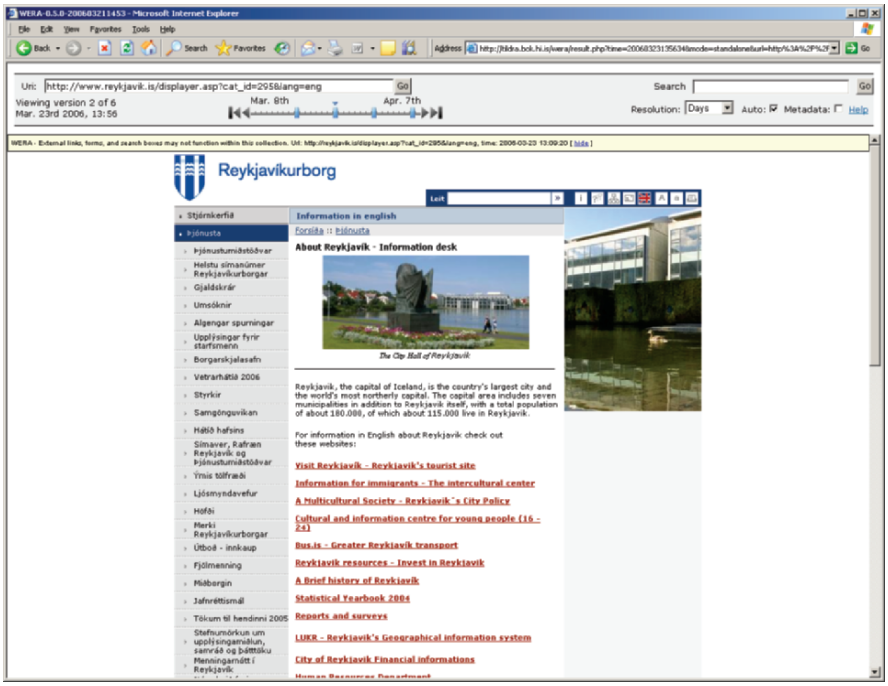


Fig. 6.5. Access to archived page using WERA with the timeline, and various informations at the top of the page

Acknowledgements

This chapter is based on published and unpublished IIPC documents and discussions in the IIPC Access Working Group.

References

All references were looked up in May 2006.

Data Mining: http://searchsqlserver.techtarget.com/sDefinition/0,290660, sid87_gci211901,00.html

Pandora: http://pandora.nla.gov.au/index.html

Minerva: http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html

META tag in HTML: http://searchwebservices.techtarget.com/sDefinition/0,290660, sid26_gci542231,00.html

Dublin Core Metadata Initiative: http://dublincore.org/

Digital Object Identifier (DOI): http://www.doi.org/

Open Archive Information Standard (OAIS): http://www.rlg.org/longterm/oais.html

Internet Archive: http://www.archive.org/

The American Society of Indexers: http://www.asindexing.org/site/webndx.shtml

PageRank™: http://www.google.com/technology/

IIPC Use Case document: http://www.netpreserve.org/publications/reports.php

Nutch: http://lucene.apache.org/nutch/docs/en/

NutchWAX: http://archive -access.sourceforge.net/projects/nutch/index.html

MD5 algorithm: http://userpages.umbc.edu/~mabzug1/cs/md5/md5.html

Sigurdsson Kristinn: Eliminating duplicate non-text documents in a snapshot crawl. Unpublished IIPC document

Wayback Machine: http://www.archive.org/web/web.php

WERA (WEb ARchive Access): http://archive-access.sourceforge.net/projects/wera/

Alexa: http://pages.alexa.com/company/history.html

NWA Toolset: http://nwa.nb.no/

7 Mining Web Collections

Andreas Aschenbrenner[1] and Andreas Rauber[2]

[1] Goettingen State and University Library
[2] Vienna University of Technology

7.1 Introduction

First contact with Web archiving is often with the technological issues in collecting Web material, which is discussed in the first part of this book. However, it is also one of the key messages of this book that Web archiving and the stewardship of Web material over the long term entails an array of tasks and functions. Thus, after the first part of this book on the more technical issues in building Web archives, this part discusses the usage of Web archives followed by their preservation in the next part. It is the ambition of this chapter to highlight the intricate interrelations between Web archive construction, usage, and preservation, to illustrate the myriad of issues involved in Web archive usage, and to convey the importance of planning and organisation of Web archives with respect to their later usage.

Usage of Web archives intuitively implies a module building on an existing Web archive that allows access in a similar manner to the way we access the current Web with the additional dimension of the time. Thorsteinn Hallgrímsson describes such a module in Chap. 6, and the popular access point to the Internet Archive, the Wayback Machine is highlighted in a respective case study in the last part of this book. Besides such access modules, however, there are a wealth of other tools and techniques for web archive usage, supporting access to web archives and the analysis of their content. Additional means for access and analysis enhance the value of a Web archive enormously, and may even attract entirely new target audiences. These means are to a great extent subsumed in the broad field "Web mining". What are the (hidden) sources in a Web archive and what usage scenarios can we infer from them? This chapter describes what "Web mining" refers to and outlines various usage scenarios based on

selected case studies that excel in unlocking the hidden information of Web archives.

The creation of novel services and the attraction of more clients create enhanced business value of the Web archives, yet they also entail additional tasks and responsibilities. Usage is more than a mere add-on module on any Web archive, and demands adequate planning and effort. Particular usage schemes may demand collecting specific Web material and may raise requirements for their management and preservation. Similarly, the needs of the archive's user group need to be taken into account from the outset. This shows the tight integration of a Web archive's usage in all its other functions including collection and preservation as well. By exploring possible usage schemes and specifying the Web material they are based on, this chapter therefore aims to support delineating a Web archive's dedicated community and its scope, defining requirements for collection and archival management, as well as for the preservation of the collections.

Moreover, exploring possible usage scenarios and selecting the appropriate ones is the basis for establishing the envisioned role of the Web archive within an organisation or for society. This, in turn, is a prerequisite for shaping the organisational structure of the archive, eliciting funding sources, and establishing requirements for the technical environment. All these points will help to ensure the archives' sustainability in the various dimensions of the word: organisational, financial, as well as technological sustainability. For all these reasons the exploration of usage mechanisms is at the core of any Web archiving initiative.

In addition to the challenges in constructing a Web archive and maintaining a successful service, a Web archive and specifically the combination of different data sources and mining techniques may have a significant social impact that needs to be considered from the outset. This regards most notably privacy concerns relating to Web archives equally as to any data collection initiative. Leaving this issue to (national) legislation falls short of resolving this ethical issue; it demands an open discussion involving all stake holders, including those responsible for the Web archive, data creators and providers, as well as the users. Without any specific and general recommendation being available at this point of time, Web archivers and Web miners should keep in mind the social repercussions of the analyses they are performing, and incorporate strategies to calm privacy concerns by anonymisation, time-delayed access, and by trying to foresee the potential consequences of the kind and amounts of data they amass – while at the same time obviously trying to keep their archive and analyses comprehensive and powerful to live up to both its scientific and social value.

Web mining is a broad field that embraces a variety of techniques for the extraction of patterns from and for the analysis of a web archive's

holdings. It is a very dynamic scientific field that grows with the proliferation of information and services on the Internet, constantly adapts to new Internet technologies, and incorporates new analysis techniques. Indeed, various attempts to structure the field according to the Web material used or the mining technique were dwarfed by the field's rapid development. An early account (Cooley et al. 1997), for example, distinguishes between Web content mining and Web usage mining. This classification differentiates along the different kinds of material used; the actual Web pages on the one hand and access log files on the other. Only in the late 1990s techniques such as the PageRank Algorithm (Brin and Page 1998) employed by the search engine Google[1] were developed, which draw on the links embedded in Web content rather than the content data itself. This development added another area of Web mining, namely Web structure mining, which later descriptions of the field accounted for. Today, techniques to analyse Internet infrastructure data and Web server transfer meta-data enhance the list of Web material apt for Web mining, and the emerging field of network analysis is a new addition to Web mining techniques. While this chapter has no ambition to develop an updated typology for Web mining, it will present an updated view on the issues in Web mining with its requirements and possibilities. This account focuses purely on Web mining use cases that are currently applied or could potentially be applied to Web archives while omitting techniques such as wrapper induction (Kushmerick 2000) or other, purely Information-Retrieval oriented methods and projects. While it is impossible to provide a full account now or in the future, this chapter provides a broad synopsis of the field and equips the reader with the necessary triggers for shaping her or his Web archiving initiative. The upcoming section provides a general overview of the various kinds of Web material that are employed by Web mining techniques either isolated from each other or in any combination. These diverse resources may be considered for inclusion in Web archives to allow for existing Web mining applications to be installed and foster research. A subsequent description of selected case studies illustrates the range of conceivable usage scenarios and potential services.

7.2 Material for Web Archives

Various kinds of data offer themselves for inclusion in a Web archive. Some of these data are very obvious; some of them less so, but they can

[1] Google. www.google.com

easily be collected alongside other activities; and some data may be available for some organisations, however, more difficult to obtain for others. The following list gives an idea of the diversity of Web material at our finger tips. These various data are employed by the use cases compiled in the "Use Cases" section later.

7.2.1 Web Pages

The first and most important material for Web archives obviously are the actual Web pages. Their collection has been exhaustively covered in this book's Chaps. 1, 3–5. Stressing again the myriad of data formats available on the world wide Web, the pioneering Swedish Web archiving initiative Kulturarw3 (Arvidson et al. 2000) found 465 different file formats in their 10th sweep of the Swedish Web space, which took almost ten month from August 2002.[2] The five most frequent file types, which covered 96% of the files captured in this sweep, are HTML (HyperText Markup Language) documents, images in GIF and JPEG format, as well as plain text and PDF documents. Other types include various sorts of multimedia and presentations.

The most characteristic file type for the world wide Web and, indeed, the most prevalent type that constitutes almost half of all the online available documents, are documents written in HTML that are interpreted by Web browser applications. Besides static Web pages, HTML documents may be dynamically generated by a Web server to satisfy a specific user request. Web pages with a file extension PHP, CGI, or the like point to dynamically generated pages. All these documents may be marked up in HTML versions up to HTML 4 or the eXtensible HyperText Markup Language XHTML.[3] HTML is the document type in the focus of most current Web mining activities. In the future, Web mining may incorporate techniques from multimedia mining (Zaiane 1999) and other data mining areas, but at this point of time the field is mainly focusing on the distinctive properties of HTML.

When dissecting the name "HyperText Markup Language" we find its first main feature to be hypertext. Hypertext has links embedded in its textual content that enable the reader to control the flow of reading by jumping from one page to another in a self-chosen order and fit to her needs.

---

[2] Kulturarw3, Long time preservation of electronic documents. The web archiving initiative of the Swedish Royal Library. http://www.kb.se/kw3/

[3] The World Wide Web Consortium (W3C) HyperText Markup Language (HTML). http://www.w3.org/MarkUp/

This non-linear way of reading has entirely changed traditional publishing paradigms. The hyperlink in one page pointing to another page also marks a relation between the two pages. Based on this relation a graph of a website or, theoretically, the whole Web can be drawn. In such a Web graph a page is a vertex, and each link within the page is an edge to another vertex. Viewing the Web as a graph offers a wholly new perspective that largely fades out the textual content of a page and focuses on the relations between Web documents.

HTML's second main feature is its semi-structured nature, as the term "markup" indicates. Thereby, each HTML document consists of a number of chunks that are started and ended by a, respective, tag. For example the highest level heading of a page is started with <H1> and terminated with </H1>. Tags may be recursively embedded within each other. Structuring the HTML document in clearly defined chunks allows automatic tools a basic level of machine comprehension. It is HTML's semi-structured nature that greatly influences and enhances the possibilities of Web mining applications.

7.2.2 Metadata

Metadata – chiefly defined as data about data – also play a pivotal role on the Web and consequently also in Web archives. These background data conveying a wealth of information in addition to the obvious content of the Web objects can be gleaned from the Web or they are produced through the application of mining techniques.

There are two main metadata types of interest with regard to Web mining: metadata about the object from the Web, as well as technical metadata in context with the transmission of the object via the Internet. Metadata about a Web object can be obtained from various sources. One major source for metadata is, of course, the object itself including its date of creation and object size. The file format can be extracted, either by simply looking at the file extension of the object or by applying tools like JHOVE[4] that identifiy the format based on internal properties of the file. The human language used in textual documents is another metadata item to be gained from the object itself by help of specialised external language detection algorithms, besides a number of other conceivable metadata items. As a further metadata source, the object may carry some limited metadata in dedicated

[4] JHOVE – JSTOR/Harvard Object Validation Environment. http://hul.harvard.edu/jhove/jhove.html

fields, such as for example Dublin Core[5] mark-up. Tools such as UKOLN's DCdot[6] can extract these metadata easily. Due to the lack of a central control in the Web, however, there is rarely consistent metadata in objects, if at all. This source will hence be only fruitful for organisations that have the authority to centrally impose a specific structure on their Web material including specific metadata. This source is only of limited use for initiatives that tackle the open Web or corporate Web mining initiatives unable to control the addition of metadata to their Web material. Lastly, communication with the Web server offers some limited but important object metadata including its location (i.e., the URL) and the MIME type of the object.

Communication with Web servers also yields technical metadata in context with the transmission of the object via the Internet. When an object is requested by a Web browser, the browser typically opens a HTTP connection to port 80 of a Web server and requests a specific document. The response from the Web server contains the requested object if available, as well as additional metadata including information about Web server software (e.g., "Apache/1.3.26 (Unix) PHP/4.3.3 mod_ssl/2.8.9 OpenSSL/-0.9.6c"), the connection status in a standardised code, and the MIME type of the transferred Web object. As we will see in the use cases later, this information gives some valuable information about the technology infrastructure in the Internet.

7.2.3 Usage Data

The primary source for usage data are server logs (Mobasher 2004). Every time a user sends a request to a Web server, the Web server protocols that together with some additional data about the request and where it came from. A standard server log format is the W3C Extended Log File Format (Hallam-Baker and Behlendorf 1996); others such as Microsoft IIS Log Files hold similar information with some minor format variations. A log file may include along with transmission data such as date of the request and number of bytes transmitted also user information such as the user IP address, user name or the computer name of the user. Web server logs are easy to get for corporate Web archiving initiatives. External Web archiving initiatives will, however, only get this data with the cooperation of the respective organisation and the respective Web server administrator.

[5] Dublin Core. www.dublincore.org

[6] DCdot – UKOLN (UK Office for Library Networking) Dublin Core metadata editor. http://www.ukoln.ac.uk/metadata/dcdot/

Other usage data may include personal user information created by a specific Web service. For example, some websites may be restricted to registered users and allow the user to create a personal user profile. Online services may record a history of past user navigation patterns, and perhaps include explicit user ratings on products as a basis of collaborative filtering systems. The Web navigation service Alexa,[7] for example, builds on user ratings and traffic rankings. These data are, however, reserved to the respective Web site hosts providing such a special service. Passing this data on to third parties may raise severe data protection concerns.

7.2.4 Infrastructure Data

As mentioned earlier, the technical metadata retrieved in the course of HTTP transmissions offers some limited information about the technology setup of Web servers. This data, however, still does not reveal information about the overall infrastructure of the Internet and how the various Web servers are interconnected. Data that reflects the Internet infrastructure are the routing tables. When data is routed through the Internet it is passed from one local network – also referred to as an autonomous system – to the neighbouring unit until it reaches the destination. Routers are the entities that forwards the data packages, and routing tables tell them which neighbouring autonomous system is next in the data package's path to its final destination. The Border Gateway Protocol (BGP) (Rekhter and Li 1995) is a protocol for exchanging routing information between routers. BGP data contains some resources that reflect Internet connectivity and the overall technology infrastructure at the Internet backbone connections, including IP addresses, autonomous system numbers, CPU cycles in routers and bandwidth consumed by routing update traffic (Brody and Hickman 2000). BGP data is, however, difficult to obtain. There are several initiatives that archive and provide BGP related traffic data. For example, the National Laboratory for Applied Network Research's Measurement and Operations Analysis Team (MOAT)[8] provides BGP data from almost 50 contributors who deposited BGP tables between November 1997 and March 2001. Other such initiatives include the Internet Traffic Archive,[9] as

[7] Alexa. http://www.alexa.com/

[8] National Laboratory for Applied Network Research (NLANR), Measurement and Network Analysis Group: Network Analysis Infrastructure (NAI). http://moat.nlanr.net/infrastructure.html

[9] ACM SIGCOMM: Internet Traffic Archive. http://ita.ee.lbl.gov/index.html

well as the Ripe NCC Routing Information Service [10] that provides a myriad of historical information about Internet traffic.

However, while BGP captures a detailed view of the centre of the Internet, the connectivity of the periphery is rarely captured. Active probing of Internet connectivity by using the Traceroute tool may yield a broader coverage, even though it fails to provide data as detailed as the BGP routing tables. Traceroute is a network administration tool capable of determining the route data packets take to reach a destination host. Tools based on the Traceroute utility can be implemented by anybody (though uncoordinated massive scanning from numerous initiatives would probably hamper global data traffic and annoy network administrators). A comprehensive archive has already been set up by the Internet Mapping Project (Cheswick and Burch, 2004) that started to collect Traceroute traces in 1998 and has the long-term mission of acquiring and saving Internet topological data.

7.3 Other Types of Information

This list of Internet data types apt for Web archiving is, of course, not exhaustive. Over time a myriad of other Web material is likely to emerge only to perish again after being superseded by the next generation data type. Already now it is hardly possible to keep track of the host of different data types on the Internet. Applications and protocols such as the IRC chat channel, Multi User Dungeons and other online games, newsgroups, RSS (Really Simple Syndication Web content and distribution), mailing lists, and many others more are excluded from the above succinct and general list for the impossibility to provide a full account. Material considered for inclusion into a Web archive may even cover extraneous sources, such as monitoring (e.g., filming) user sessions within specialized user study labs, to capture the interactive nature of some Web applications, online games and other highly interactive elements on the Web.

Also, the accessibility of specific data may vary between Web archiving initiatives. Each Web archive therefore needs to identify for itself the data types, the scope of collection and quality requirements to fulfil its mission and to best serve its dedicated community. Some archives may choose to specialise on a single, perhaps one of the more rare data types. The Netscan Project,[11] which maps the social geography of usenet news, is exemplary for the valuable service and exciting research such a specialised archive

[10] Réseaux IP Européens (RIPE): Routing Information Service (RIS).
[11] Netscan Project. http://netscan.research.microsoft.com/

may provide. Other initiatives may choose to collect a variety of different data types, for a variety of different services or perhaps in anticipation of possibilities for future usage. In the latter case, the regular environmental scan by the National Library of Australia (Philips 2003) may be a model for other initiatives struggling to keep their corporate selection and collection policy updated in the rapidly changing environment.

As the use cases in the next section describe, Web mining techniques mostly need a very particular data chunk from the various Web material described earlier. Some techniques combine a variety of different Web material, yet they focus on specific data elements such as the name of the Web host in an external Web link. Web mining usually involves a preprocessing stage, in which the original data is tailored to the specific needs of the mining technique. In this stage specific data may be converted or augmented with other, external data, and other data may be dumped as it is not useful in the particular case. While these processed data are all needed for a specific data mining technique, the original material needs to be archived as well. Without the original data an initiative is confined to a small set of mining techniques. Moreover, perhaps a new version of a mining technique demands additional data elements. A Web mining initiative should therefore avoid obstructing the evolution of mining techniques and the development of the archive, and rather preserve the original data.

In addition to the original data, the context of the data to be preserved in a Web archive needs to be captured in many cases as well. Without exhaustive documentation future generations will hardly be able to make sense of the material. For infrastructure data, for example, documentation of the local host configuration is needed, and also of the surrounding Web environment (e.g., prevalent applications and prevailing protocols at that time). As another example, in the case of usage data real-world promotional activities may impact hugely on the user access rate and consequently the Web logs, so even a record of an organisation's real-world activities could support future interpretation of the material collected by a Web archiving initiative. The preservation of context data alongside the original data could, of course, escalate easily and become unfeasible, yet it should be considered and clearly defined by the Web archiving initiative when establishing its collection policy.

7.4 Use Cases

The different kinds of data listed in the last chapter form the basis for a variety of applications. Some initiatives building on such data from Web archives

are presented in the following. They feature both initiatives that focus solely on one type of data and others utilising a combination of them. A combination of different data types allows Web archives to serve a larger range of user communities and facilitates the application of entirely new, perhaps more reliable, mining techniques.

The following use cases are selected to convey the scale of inherently different mining techniques that may be built on Web archives. They highlight the valuable services a Web archive may fulfil. Indeed, Web mining techniques can unveil valuable information pertaining to all walks of life. Future Web archives may provide indispensable services for research, business and private life in all conceivable areas from research in sociology and mathematical theory to marketing analysis and public opinion tickers.

7.4.1 Analysing Web Content and Web Technology

With respect to the usage of Web archives, several projects are developing access interfaces that allow users to search and surf within such an archive, such as, e.g., the Internet Archive's Wayback Machine, or the access tools for the Nordic Web Archive (NWA). These tools provide for each URL a timeline listing the dates when this specific URL was added to the archive, i.e., which versions of the respective file are available.

Going beyond the mere navigation within the archive as a mirror of the World Wide Web existing at the respective times, several projects take a more structured approach to storing and analysing the Web. The Web Archaeology project (Leung et al. 2001) studies the content of the World Wide Web using a variety of content representations, referred to as features, including *links* capturing connectivity, *shingleprints* capturing syntactic similarities, and *term vectors* capturing semantic similarities. The Mercator Extensible Web Crawler is used for large-scale data acquisition, and specific database models were developed at the second layer of the system architecture for storing the feature databases. Various tools are added to the top layer of the system architecture to facilitate specific types of analysis, such as, e.g., in the Geodesy project trying to discover and measure the structure of the Web.

Another Web page repository is being built within the WebBase project at Stanford University, addressing issues such as the functional design, storage management, as well as indexing modules for Web repositories (Hirai 2000). The main goal of this project is to acquire and store locally a subset of a given Web space in order to facilitate the performance execution of several types of analyses and queries, such as page ranking, and

information retrieval. However, it limits its scope to the archiving of one copy of each page at a time, thus providing no historisation, and focuses on HTML pages only.

When it comes to analysing large amounts of data in a flexible manner, data warehouses (DWH) have evolved into the core components of decision support systems (Kimball, 2002). A DWH is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of decision-making processes (Inmon 1992). Rather than storing data with respect to a specific application, the information is processed for analytical purposes, allowing it to be viewed from different perspectives in an interactive manner. It furthermore integrates information from a variety of sources, thus enriching the data and broadening the context and value of the information.

The primary concept of a DWH is the separation of information into two main categories, referred to as facts and dimensions, respectively. Facts are the information that is to be analysed, with respect to its dimensions, which often reflect business perspectives, such as a geographic location, evolution over time, product groups, merchandising campaigns, or stock maintenance. The DWH may be envisioned as a multi-dimensional data cube. Using on-line analytical processing (OLAP) tools, this data cube allows to interactively drill-down, roll-up, slice and dice, view and analyse the data from different perspectives, and to derive ratios and compute measures across many dimensions. These OLAP operations assist in interactive and fast retrieval of 2D and 3D cross-tables and chart-table data from the cube, which allow convenient querying and analysis of a Web data storage.

This technology is used to analyse web collection data in the WHOWEDA project, pursued by the Web Warehousing and Data Mining Group at the Nanyang Technological University in Singapore (Bhowmick et al. 2000). Within this project a DWH stores consecutive versions of Web pages, adding a time dimension to the analysis of content and link structure. URL, size, date of last modification (and validity, with respect to subsequent visits to a given site), size, etc. are stored together with the content and structure of a document. Furthermore, link information, as well as the position of links within documents are recorded and made available for further analysis. Although a more structured approach to the analysis of Web pages is taken within the scope of this project, it primarily focuses on a detailed analysis and representation of the content of the documents.

Technologically similar, the Austrian on-line archive processing project (AOLAP) (Rauber et al. 2002) uses DWH technology to analyse the data acquired by the Austrian on-line archive (AOLA) (Rauber and Aschenbrenner 2001). The archive consists of Web pages, including all

types of files as collected by the harvesting software, and rests on tape archives organised primarily according to domain names. In addition to the actual pages, meta-information that is provided or created during the crawling process, is collected and stored as part of the archived files. This includes information provided as part of the HTTP protocol as well as other information provided by the server, such as the server software type and version, the operating system used by the server, date and time settings at the server, as well as last-modified dates for the respective file being downloaded.

The information extracted from the pages includes *file types* based on file extensions, *file size*, internal and external *links*, information about *frames*, *e-mail addresses* and interactive *forms* used in the case of HTML files, *date of last modification*, and others. With respect to the various domains it mainly concentrates on *IP addresses* and thus *network types*, *operating system* and *Web server software* information. Furthermore, AOLAP integrates information from other sources, to enrich the data provided by the harvesting system. Specifically, a set of WHOIS servers is used to provide geographic location information of Web service registrars, alias names, etc. The information is further transformed and loaded into a relational DBMS using a star-model like design for the data storage. Based on this model a multi-dimensional cube is created which can further be used for interactive analysis, such as the distribution of file types across different Web servers or link structure analysis.

The number of file types encountered in the Web archive is highly relevant with respect to the preservation of the archive that is, keeping the pages viewable in the near and far future. The number of types also represents a good mirror of the diversity of the Web with respect to the technologies employed for conveying information. Overall, AOLAP encountered more than 200,000 different types of files based on their extensions, and more than 200 different types of information representation when using the MIME type as the indicative criterion. However, the quality of the information provided this way is rather low, as a large number of both file extensions as well as MIME types are actually invalid, such as files with extensions .htmo, .chtml or .median, .documentation. While the majority of file extensions encountered definitely are erroneous, these are indicators of serious problems with respect to preserving that kind of information, as well as the need to define solutions for cleaning this dimension to obtain correct content type descriptors.

Analysing the link structure confirms an intuitive tendency, namely a high inter-linkage within each respective domain, i.e., .com sites linking mostly to other .com sites, .cc linking within .cc and so on. However, there

are also some interesting exceptions to this rule, which can be interactively analysed by drilling down in the respective domains.

Since AOLA collects a huge range of metainformation during the crawl, these various types of information can be put in relation to each other, such as for example the distribution of different Web server types across the various domains, and the evolution of market shares of Web servers.

Obviously, further types of information can be extracted from the Web pages and integrated into a DWH (e.g., automatic language detection methods) covering in larger detail additional technological information, such as the usage of cookies, embedded Java applets, Flash plug-ins, encryption, and others. Furthermore, being able to analyse the content-based dimension of a Web archive provides the basis for subject gateways on a variety of topics and with a historic dimension.

Due to the flexibility offered by the DWH-based approach, these systems can be used in a wide range of Web archive utilisation scenarios, both for archive maintenance, as well as for exploiting the information constituted by the archive. The information obtainable via these DWH-based approaches is similar to statistics computed by some of the longer-running projects in this field, such as the Swedish Kulturarw3 project, which has available data from several complete runs since 1996. These statistics show, for example, the first traces of XML documents in early 1999 and reveal how XML documents have been increasing in number and as a share of Web documents available from 1999 to date. Another fascinating example is the surprisingly sudden victory of PDF over the Postscript file-format within about a year in 1998.

The main benefit of the proposed DWH approach for Web archive analysis lies in the flexibility with which interactive analysis of the archive can be performed. Contrary to most current approaches, the focus of this type of analysis is not primarily on the content of the data, but rather on meta-information about the data, as well as data about the technologies used to provide a given service.

7.4.2 Exploring Web Communities

Within a minimal time-span the Web has become the communication and publishing space of modern society. It grew from an experimentation space for scientists and computer geeks in its early times, to the all-inclusive medium nowadays. With all walks of life represented on the Web, distinct communities can be identified that form a Web of personal relationship and focus on a specific topic or theme. Different communities may include fans of a music group or supporters of a sports club, who shout out their

fondness on the web; fond users of a service, say the online catalogue of a public library, who link to the library's website; or also the players of a specific online game.

Various initiatives are researching methods for automatically identifying Web communities, and their motivations are manifold (Kumar 1999). For instance, for improving search-engines and creating new services, knowledge of Web communities and their structural characteristics is of eminent importance. Where search engines used to be praised for their comprehensiveness, returning a rich set of more or less relevant Web pages to a specific query, Web users are increasingly overwhelmed by the flood of information available on the Web. The search engine Google excelled to become the number one search engine not because it returned the most results of all the search engines, but because it ranked the results according to their estimated value and thereby greatly assisted the user in finding relevant information. The technique for identifying Web communities may allow the next quantum leap in search technology. A user may then be able to better focus on a specific area of interest and prune all the other information (Flake et al. 2002). Visionary installations of this technique include advanced content filtering and the largely automatic creation of web portals, ensuring that they are always up-to-date and as complete as possible. Recommendation systems for specialised fields such as the music recommender by (Knees et al. 2005) are within reach.

Identifying Web communities enables producing opinion polls to any kind of topic at any time. The evolution of opinions can be tracked and comparisons established. One of the currently most vibrant online community spaces, the body of public weblogs, illustrates this opportunity. In the years from 1999 onwards, the public blogosphere grew at a tremendous rate. Soon a number of blog engines emerged to provide search and a variety of other services based on mining techniques. For instance Daypop[12] offers "Word Bursts" of often used words and phrases, and Technorati features ranking services services on blogs, news, books, and movies. BlogPulse[13] provides the exciting services "Trend Search" and "Conversation Tracker". The latter extracts sequences of blogs, comments and re-blogs in a single conversational trail. "Trend Search" gives the blogosphere's level of interest into a topic as it develops over time (Aschenbrenner and Miksch 2005).

By 2005 the social network of public weblogs continues its highly dynamic development. New features pop up faster than those existing can be explored, and the blogosphere is increasingly interwoven with other

[12] Daypop. www.daypop.com
[13] BlogPulse. www.blogpulse.com

emerging applications of the social software movement. More exciting applications of mining techniques can be expected from the blogosphere community, which is a great inspiration for respective mining activities in Web archives.

Mining specific communities and the evolution of the communities themselves also reflects public trends that can be used in business. Companies may choose to align their advertising campaigns according to these trends, or they may direct online advertising campaigns to specific communities. "Community pressure" may even lead to abandoning a specific product, or indicate market gaps. All in all a company's strategic management could be informed by various conceivable analyses building on the identification of Web communities (Reid 2003).

Web communities are particularly dynamic social entities that may emerge only to quickly vanish again, or they may persist over extended periods of time. Web archives that record this offer a valuable resource for sociological research. These archives can answer questions such as the effect of a historical event on society: How many disparate communities formed around a specific event (e.g., a sports event, an election, or a political topic)? How large did they grow? and how long did they last?

Various techniques have been developed to identify Web communities. All these techniques build on the hyperlink structure of the Web, assuming that the members of the same community tend to reference more often their peers than external Web pages. In other words, identification of Web communities comes down to clustering more tightly coupled Web pages or websites together. Techniques include Bibliographic Metrics and Bipartite Cores, which build a community starting from a local set of Web pages; the HITS (hyperlink-induced topic search) Communities and the PageRank Communities algorithms, which consider all edges in the global Web graph for building a community; and the Community Algorithm, which equally works on the global Web graph as on a local sub-graph (Flake et al. 2003). The data used for community identification are basically only the links in the content of the Web pages. Specific techniques may use the text surrounding the links as a supplementary factor, but essentially no additional data than the plain Web content is used.

Experiences in community extraction show that the number of cyber-communities has consistently increased over time (Donato et al. 2004). Toyoda and Kitsuregawa analysed the evolution of Web communities in a Web archive of Japanese websites with annual crawls starting from 1999 to 2002 (Toyoda and Kitsuregawa 2003). They found that communities are relatively stable, considerably more stable than the stability of Web pages. However, although the communities as such are relatively stable, their structure changes dynamically; a majority of the identified communities is

involved in merges and splits from one snapshot in the Web archive to the subsequent one. While the number of Web communities more than doubled between 1999 and 2002, the size of the individual communities was stable over that time.

These initial experiments already show the intriguing findings and the multiplicity of applications that community identification on public Web pages can produce. As the necessary data will be available in most Web archives, this branch of Web mining is a valuable addition to all Web archiving initiatives.

7.4.3 Screening Web Users

Web usage mining is increasingly employed by corporate websites and in eCommerce. It mainly builds on Web usage. On a basic level it answers questions like what parts of a website a user visited in a single session, and where she spent most time. Starting from such basic information, a number of valuable analyses can be conducted for organisational websites. Web log analysis is used for characterising the designated community of an online service (Brody and Hickman 2000), based on usage patterns the online effect of promotional activities can be gauged, or the structure of the site map can be improved. For instance, if repeatedly users view a combination of Web pages in a single session it may be convenient to provide a direct hyperlink to and from those pages. The website may even be adaptive in the sense that its Web pages adapt themselves dynamically to improve site navigation (Perkowitz and Etzioni 1997).

Usage mining can also be employed for recommender systems (Mobasher 2004) In fact, recommender system are increasingly wide spread in online shops, and many of us are likely to have encountered the kind suggestion “Users who viewed/bought this product also viewed ...”. While recommender systems mostly work with anonymous users, personalisation of services is the current buzzword that promises to revolutionise Web navigation. Thereby, user data are stored at the Web server or in cookies on the client side and reused at later sessions. It is, in fact, possible to infer various demographic data with a relatively high accuracy only based on user session data, including gender, age, income, and marital status (Murray and Durrell 2000). Such personal information may inform business policy making or may facilitate more directed advertising.

While usage mining is an increasingly important eCommerce analysis activity, there is at this time no apparent long-term value of usage data. It may be a valuable source for sociological research in the far future, but clear usage scenarios are still missing. For comparison, the longitudinal

study of Web usage (Cothey 2002) took a view over a ten-month period, where longitudinal studies in other areas including Web content usually take a view over years and decades. Only (Covey 2002) takes a slightly more long-term view and also underlines the value of mining user logs over a number of different websites and organisations. Only just the fact that the value of usage data is unclear at this point of time does, however, not imply that it is worthless. This still is a young field and scholars in the future may find usage data a truly rich source.

Thus, while the long-term value of user logs may not be obvious to Web archiving initiatives tackling the broad collection of external websites today, they still bear a unique type of information that may prove indispensable in the future. Web usage data may, for example, facilitate the identification of relevant sub-trees of a website, and may thereby solve automatically what the Pandora[14] project at the National Library of Australia approaches in a manual way. Also (Chakrabarti et al. 2000) points to the value of personal references created by users intentionally or as a side product of Web browsing. Community powered Web archives such as the DACHS archive (Gross 2003) already now have the possibility for preserving this kind of data. Also community services as the one provided by Alexa[15] are in position to preserve user session data, and may inform the collection policies of Web archives or serve as a prolific data source for research.

7.4.4 Researching Networks

The previously described use cases became gradually more abstract, starting from surveying the actual content and Web usage, to analysing who visits websites. This section adds another level of abstraction and analyses the Web as an entity. The Internet is a prolific place for theoretical and mathematical research. Current research on the Internet includes self-organisation and fractal growth, graph theory, as well as game-theoretic analyses (Czumaj et al. 2002).

Probably the most exciting area in this respect currently is network theory. In fact, the Web was the trigger for a quantum leap in network theory (Bharat et al. 2003). In the late 1990s researchers found many properties of the Web to be incompatible with traditional network theory. Complex networks were formerly considered to be completely random. The link structure of the Web, however, has a considerable number of nodes, to which an

[14] Pandora project at the National Library of Australia. http://pandora.nla.gov.au/
[15] Alexa. http://www.alexa.com/

enormously high number of links point to. These nodes are called "hubs" and they reflect a property of the Web called "scale free", which basically means that already popular Web pages tend to become even more popular on the Web. Properties of the Web that allow the Web to be scale-free in the first place are its growing and dynamic nature. This was not considered by network theory until the late 1990s.

Since the discovery that the Internet defies traditional paradigms, an array of other scale-free networks has been discovered in diverse scientific areas (Albert and Barabási 2002). They include protein interaction maps in cells, social relationships, research collaborations and global trade networks. The research assumption is that the underlying mechanisms as discovered in the Web apply equally to all complex, dynamic networks, and that any findings can be transferred between these seemingly so different networks. Therefore, a number of professions including sociologists, biologists, and computer scientists follow current developments in network research with equal interest.

Currently, various initiatives are embarking on massive research efforts regarding complex networks, including the research group at the University of Notre Dame[16] and the European Commission funded project COSIN – COevolution and Self-organisation In dynamical Networks.[17] Their goal is to improve the stability, efficiency and functionality of artificial complex networks such as the Web. While dynamical networks are robust against accidental failure they are vulnerable to coordinated attacks, since the corruption of a few central nodes may disturb the whole system. Dynamical network theory may, therefore, enable new approaches to remediate hacker and virus attacks. Also, biological viruses spread in scale-free human networks, and further research on epidemic spreading may guide immunization policies in the future. A myriad of other theories and applications may emerge from dynamical networks in areas including group dynamics, linguistics, and crisis prevention in global markets. Much remains to be done until applications as mentioned earlier will materialise.

Recent research (Pennock et al. 2002) indicates that while the Web is scale-free on a high level this does not apply at a lower level; in focused communities and between peers the link distribution is less biased and not scale-free. Also the dynamic evolution of complex networks still poses many open questions.

---

[16] Center for Complex Network Research, University of Notre Dame. www.nd.edu/~networks

[17] COSIN – COevolution and Self-organisation In dynamical Networks. IST-2001-33555. http://www.cosin.org/

The data used for network theory research includes both Web pages as well as infrastructure data. Web archives could be the test bed for empirical research and experiments on the numerous open questions of a young and prosperous field of science.

7.4.5 Planning Network Infrastructure

As the Internet is the hinge of the modern communications infrastructure, its smooth operations are essential for business, government, and the public. Various organisations and initiatives have embarked on monitoring and analysing global network traffic, to ultimately make networks more robust and efficient. Amongst those are the Internet performance measurement and analysis (IPMA),[18] the CNRG Research Group on Internet Infrastructure,[19] the European Project SCAMPI,[20] as well as the cooperative association for internet data analysis (CAIDA).[21] Their findings and experiences focus on identifying network traffic congestion areas and ensuring that consumption is evenly spread between existing resources, as well as on projecting future network traffic development and simulating the impact of novel protocols and applications. These analyses are the basis for optimal deployment and configuration of existing network resources, and for guiding future investments in network infrastructure on both a global and a corporate level.

An example for the application of network analysis is Lumeta Corporation[22] that uses traceroute probes to explore corporate networks. Thereby, they are capable of identifying leaks or misconfigurations in the network. They can even identify and visualise organisational changes that impact on the corporate network. After mergers and acquisitions, for example, formerly separate corporate networks grow together, which can be followed in an archive containing traceroute data of the respective networks. Lumeta is a spin-off company from the Internet mapping project (Cheswick and Burch 2004) described earlier, and it continues to maintain and extend the archive of daily Internet traceroute probes that was started by the Internet Mapping Project in 1998. This database is used by various initiatives to study Internet evolution, as well as diverse other matters including graph

[18] Internet Performance Measurement and Analysis (IPMA). http://www.merit.edu/~ipma/

[19] CNRG Research Group on Internet Infrastructure. http://www.cs.cornell.edu/cnrg/

[20] SCAMPI – Scaleable Monitoring Platform for the Internet. http://www.ist-scampi.org/

[21] Cooperative Association for Internet Data Analysis (CAIDA), San Diego Supercomputing Center (SDSC). http://www.caida.org/

[22] Lumeta Corporation. http://www.lumeta.com/

theory. Indicative visualisations can be found on the archive website (Cheswick and Burch 2004), including a depiction of the changing topology of the Serbian network during the war in Yugoslavia in 1999. While it is obvious that the network took some damage during the war and some vital connections were eradicated, this time-lapse motion view shows the stability of Internet infrastructure even in face of major catastrophes.

Network analysis can this way guide policy-making and investment decisions on a corporate, national, as well as a global level. An analysis of global Internet infrastructure evolution shows that the Internet hype in the 1990s led to the installation of a massive overcapacity. Many companies failed and dropped off the Internet when the Internet bubble burst in late 2000 with continued deflation throughout 2001. The Internet economy slump can be clearly recognised in infrastructure data from that time.[23] Today the Internet infrastructure grows at a healthy rate. The rapid growth in the 1990s has stopped, however the overcapacity installed in the 1990s is still not fully used (Odlyzko 2003). It primarily grows at the periphery and thereby facilitates Internet access for a growing number of users.

Still little is known about patterns in Internet growth, but advanced research activities are working on filling the gaps. Increasingly initiatives manage to successfully combine existing approaches and knowledge Siganos et al. (2002) indicate the potential of combining various data at different levels – the network, the node, and the routing level. Also, the application of approaches from mathematical network analysis as described earlier yield a more accurate picture of infrastructure dynamics and its evolution (Vázquez et al. 2002; Siganos et al. 2002). Only a comprehensive Web archive that includes various forms of data and allows for the flexible plugging-in of novel mining techniques is able to support this kind of future research.

7.5 Conclusion

There is a myriad of conceivable Web mining applications that may be suitable for Web archives. It is the goal of this chapter to convey the breadth of the field and the opportunities at hand in order to trigger thought and pave the way for planning and design at the individual Web archiving initiative. Web archiving continues to be a highly dynamic field, and exciting new applications continue to pop up.

[23] (Broido et al. 2002), specifically Chap. 7.2.

Web archiving goes beyond technological issues and the mere collection of material from the Internet. Usage related issues need to be viewed from an organisational, functional as well as a technological perspective, and usage plays into all other functional elements of a Web archiving initiative from selection and collection, to information management and preservation. However, thorough planning and design regarding Web archive usage with a projection of possible future usage patterns by the designated community can only be done with an understanding for what is possible already now in the way of Web mining and archive usage, and an idea of the vast opportunities. Future applications need to be supported already now by an inclusive resource selection and an open system design of the Web archive.

This chapter first introduced some of the core types of data relevant for Web archiving projects. This list can no doubt be extended with more specific types and those types confined to specific initiatives for their limited availability. A full account of the possibilities and the respective requirements is impossible in the face of the diversity and the ongoing rapid development of the Internet.

Subsequently the chapter gave an impression of the possibilities in Web mining and the vastness of the field by presenting a range of research activities and Web mining projects. With some imagination it is possible to conceive of other applications future archives may provide. It has already been alluded to possible services for businesses that provide strategic data for advertisement campaign, identify emerging markets, gauge the potential for new products and thereby indicate investment opportunities. Sociology research in Web archives may identify and mitigate the digital divide with regard to geography and the varying technological opportunities and network infrastructure between regions, as well as on a societal level with mining techniques that explore the possible exclusion from the Internet of specific groups of society. As a more popular service, a Web archive may extract the "topic of the year" in the Internet and compare it with the relative coverage in traditional print media. After all, a talked-about topic in the Internet that spawns a variety of Web communities reflects the interest and opinion of the people in a more direct manner than print media.

Such opportunities trigger the vision of a future with a multiplicity of Web archive services that permeate all areas of society. Due to the still unclear requirements of such future Web mining applications, however, it is a pivotal requirement for any Web archive system to be flexible for accommodating novel mining techniques as well as new data sources. Already now Web archives should attempt to include a variety of different data as a basis for multiple, authoritative services.

As the technology environment around us changes, new forms of data will emerge that may be a valuable addition to a Web archive or that may entirely transform a Web archive. With the advent of the Semantic Web, Web archiving initiatives face novel challenges. Future technologies such as Web services and intelligent agents will be difficult to seize, even more difficult than the challenge by the deep Web as perceived today. With new approaches and tools to emerge, however, we will be able to capture aspects of the Semantic Web to the level required by the respective initiative and supported by its sphere of influence. An archive of such a semantically augmented Web sphere may offer new opportunities for Web mining and usage scenarios unheard of. An exciting future for the Internet as well as for Web archiving is wide open.

References

Aschenbrenner, A. & Miksch, S. (2005). *Blog Mining in a Corporate Environment*. Smart Agent Technologies, Research Studio. Technical Report

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74*

Arvidson, A., Persson, K. & Mannerheim, J. (2000). *The Kulturarw3 project - The Royal Swedish Web Archiw3e – An example of "complete" collection of web pages.* Paper presented at the 66th IFLA – International Federation of Library Associations and Institutions, Jerusalem

Barabási, A.-L. & Bonabeau, E. (2003). Scale-free networks. *Scientific American. 288*

Rekhter, Y. & Li, T. (1995). A Border Gateway Protocol 4 (BGP-4). *RFC 1771*

Bharat, K., Chang, B.-W., Henzinger, M. & Ruhl, M. (2001). *Who Links to Whom: Mining Linkage Between Web Sites*. Paper presented at the IEEE International Conference on Data Mining (ICDM'01), San Jose, California

Bhowmick, S., Keong, N. & Madria, S. (2000). *Web Schemas in WHOWEDA*. Paper presented at the ACM 3rd International Workshop on Data Warehousing and OLAP, Washington, DC

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems. 30*(1–7)

Brody, T. & Hickman, I. (2000). *Bibliometric Analysis: Mining the Social Life of an ePrint Archive*. The Open Citation Project: User studies: mining Web logs and user surveys. http://opcit.eprints.org/ijh198/

Broido, A., Nemeth, E., & Claffy, K. C., (2002). Internet Expansion, Refinement, and Churn. *European Transactions on Telecommunications 13*

Dodge, M. (2004). An atlas of cyberspace. http://www.cybergeography.com/

Chakrabarti, S., Srivastava, S., Subramanyam, M. & Tiwari, M. (2000). *Using Memex to Archive and Mine Community Web Browsing Experience*. Paper presented at the 9th International World Wide Web Conferenc*e*, Amsterdam.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). *Web Mining: Information and Pattern Discovery on the World Wide Web*. Paper presented at the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, CA

Cothey, V. (2002). A longitudinal study of World Wide Web users' information-searching behavior. *Journal of the American Society for Information Science and Technology 53*(2). ISSN 1532-2882

Covey, D. T. (2002). Usage and Usability Assessment: Library Practices and Concerns. *CLIR Publication 105*. Digital Library Federation, Washington

Czumaj, A., Krysta, P., & Vöcking, B. (2002). Selfish Traffic Allocation for Server Farms. Paper presented at the 34th Annual ACM Symposium on Theory of Computing, Montreal, Canada

Donato, D., Laura, L., Leonardi, S., & Millozzi, S. (2004). Large scale properties of the Webgraph. *European Journal of Physics B. 38*

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-Organization and Identification of Web Communities. *IEEE Computer 35*(3)

Flake, G. W., Tsioutsiouliklis, K., & Zhukov, L.(2003). *Methods for Mining Web Communities: Bibliometric, Spectral, and Flow*. In Poulovassilis, A., Levene, M. (Eds.), *Web Dynamics*. Springer, Berlin Heidelberg New York

Gross, J. (2003). *Learning by Doing: the Digital Archive for Chinese Studies (DACHS)*. Paper presented at the 3rd ECDL Workshop on Web Archives. Trondheim, Norway

Hirai, J., Raghavan, S., Garcia-Molina, H., & Paepcke, A. (2000). *Webbase: A Repository of Web Pages*. Paper presented at the 9th International World Wide Web Conference (WWW9). Amsterdam, The Netherlands. Elsevier Science

Cheswick, B. & Burch H. (2004). *Lumeta Corp.: Internet Mapping Project*. http://research.lumeta.com/ches/db/

Hallam-Baker, P. M. & Behlendorf, B. (1996). *Extended Log File Format*. W3C Working Draft, WD-logfile-960323

Inmon, W. (1992). *Building the Data Warehouse*. Wiley, New York

Kimball, R. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* Wiley, New York

Knees, P., Pampalk, E., & Widmer, G. (2005). Automatic Classification of Musical Artists based on Web-Data. *ÖGAI Journal 24*(1)

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for Emerging Cyber-Communities. *Computer Networks 31*(11)

Kushmerick, N. (2000). Wrapper Induction: Efficiency and expressiveness. *Artificial Intelligence 118*(1–2)

Leung, S., Perl, S., Stata, R., & Wiener, J. (2001). *Towards Web-scale Web Archaeology*. Research Report 174. Compaq Systems Research Center, Palo Alto, CA

Mobasher, B. (2004). Web Usage Mining and Personalization. In Singh, P.M. (Ed.), *Practical Handbook of Internet Computing*. CRC, West Palm Beach, FL, USA

Murray, D. & Durrell, K. (2000). Inferring demographic attributes of anonymous Internet users. *Lecture Notes in Artificial Intelligence 1836*. Springer, Berlin Heidelberg New York

Odlyzko, A. M. (2003). Internet traffic growth: Sources and implications. In Dingel, B., Weiershausen, W., Dutta, A. K., Sato, K.-I. (Eds.), *Optical Transmission Systems and Equipment for WDM (Wavelength-Division Multiplexing) Networking II*. SPIE (The International Society for Optical Engineering), 5247

Pennock, D., Flake, G., Lawrence, S., Glover, E., & Lee Giles, C. (2002). *Winners Don't Take All: Characterizing the Competition for Links on the Web*. Proceedings of the National Academy of Sciences *99*(8)

Perkowitz, M. & Etzioni, O. (1997). *Adaptive Sites: Automatically Learning from user Access Patterns*. Paper presented at the 6th International World Wide Web Conference, Santa Clara, CA

Phillips, M. (2003). Balanced Scorecard Initiative 49 - Collecting Australian Online Publications. Version 6. National Library of Australia

Rauber, A. & Aschenbrenner, A. (2001). Part of Our Culture is Born Digital - On Efforts to Preserve it for Future Generations. *TRANS. On-line Journal for Cultural Studies (Internet-Zeitschrift für Kulturwissenschaften) 10*. INST

Rauber, A., Aschenbrenner, A., & Witvoet, O. (2002). *Austrian On-Line Archive Processing: Analyzing Archives of the World Wide Web*. Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002), Rome, Italy. Springer, Berlin Heidelberg New York

Reid, E. (2003). *Identifying a Company's Non-Customer Online Communities: a Proto-typology*. Paper presented at the IEEE Hawaiian International Conference On System Sciences (HICSS 2003), Big Island, Hawaii

Siganos, G., Faloutsos, M., & Faloutsos, C. (2002). The Evolution of the Internet: Topology and Routing. *Technical Report 65*. Carnegie Mellon University, Department of Computer Science

Toyoda, M. & Kitsuregawa, M. (2003). *Extracting Evolution of Web Communities from a Series of Web Archives*. Paper presented at the 14th ACM conference on Hypertext and Hypermedia, Nottingham, UK. ACM, New York

Vázquez, A., Pastor-Satorras, R., & Vespignani, A. (2002). Large-scale topological and dynamical properties of the Internet. *Physical Review E65(066130)*, American Physical Society

Zaiane, O. R. (1999). Resource and Knowledge Discovery from the Internet and Multimedia Repositories. PhD thesis (Simon Fraser University)

8 The Long-Term Preservation of Web Content

Michael Day

UKOLN, University of Bath
m.day@ukoln.ac.uk

8.1 Introduction

Web archiving initiatives exist to collect ephemeral Web content for use by current and future generations of users. To date, most such initiatives have concentrated on the development of strategies and software tools for the collection of Web content and for providing current access to this content through interfaces like the Internet Archive's Wayback Machine. The International Internet Preservation Consortium (IIPC) is currently building on this legacy with the collaborative development of a set of tools that can be used for the capture of websites and for the navigation and searching of Web archives. The focus on collection strategies and tools is a response to what is perhaps the most significant challenge of the Web from an information management perspective. Its dynamic nature means that pages, sites and even whole domains are continually evolving or disappearing.

It is difficult to get accurate and up-to-date statistics on Web page longevity, but a range of studies hint at the ultra dynamic nature of the Web. A study by (Lawrence et al. 2001) cited an Alexa Internet estimate that pages disappeared on average after 75 days. Longitudinal studies of Web page persistence by Koehler (2004) found that just 33.8% of a sample set of pages selected in December 1996 persisted at their original URLs by May 2003. Studies of the longevity of Web references in scientific journals show similar trends. For example, a 2003 study of Internet citations in three major scientific and medical journals revealed that 27 months after publication, the proportion of inactive links rose to 13% (Dellavalle et al. 2003). The exact proportions differ, but similar results have been noted for Web-citations in other biomedical journals (Hester et al. 2004; Crichlow et al. 2004; Wren 2004), in computer science journals and conferences (Spinellis 2003; Selitto 2005), and the informetrics sub-discipline of information science (Bar-Ilan and Peritz 2004).

Web archiving initiatives deal with the ephemeral nature of the Web by harvesting selected domains or sites, thereby creating surrogates that can be used for current and future access. Current access, where this is legally possible, can be provided through initiatives own websites, as with the National Library of Australia's PANDORA archive, or through specific interfaces like the Wayback Machine, the Nordic Web Archive's access toolkit (Brygfjeld 2002), or the WERA (Web archive Access) viewer being developed by the IIPC. Longer term access, however, will depend upon initiatives being able to preserve the Web content that has been collected, thus bringing us into the domain of digital preservation.

The remainder of this chapter will outline the challenges of digital preservation, focusing in more detail on repository systems, preservation strategies and metadata. A final section will consider some specific preservation issues as they relate to the content gathered by Web archiving initiatives.

8.2 The Challenge of Long-Term Digital Preservation

Digital preservation can be understood as referring to the range of activities required to ensure that digital objects remain accessible for as long as they are needed. In a popular definition, Hedstrom (1998, p. 190) says that digital preservation involves "the planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable." Despite the growing ubiquity of digital information, the long-term preservation of information in digital form is far from a simple task. At the heart of the problem is the rapid obsolescence of the various technologies on which digital information depends, as outlined in the influential report of a task force set up in 1994 by the Commission on Preservation and Access and the Research Libraries Group (Garrett and Waters 1996, p. 2):

> Rapid changes in the means of recording information, in the formats for storage, and in the technologies for use threaten to render the life of information in the digital age as, to borrow a phrase from Hobbes, "nasty, brutish and short."

As Hedstrom's definition suggests, the challenges of digital preservation are multifaceted, involving a mixture of technical and organisation issues. While most of the underlying difficulties relate to technology, successful solutions to the digital preservation problem will depend upon what Smith (2003, p. 2) describes as the "series of actions that individuals and institutions

take to ensure that a given resource will be accessible for use at some unknown time."

The following sections will outline in slightly more detail the some of the reasons why digital information is difficult to preserve.

8.2.1 Technological Challenges

One fundamental problem is the stability of the media that digital information is stored on. A comparative study of media types undertaken in the mid-1990s suggested that, given ideal storage conditions, magnetic tapes could only reliably retain data for around 20 years, while more traditional media like acid-free paper or silver-halide microform could last for centuries (Van Bogart 1995). More recently developed storage media types may be more robust, but it is difficult to be certain about this. As Ross and Gow (1999, p. 2) have noted, "it often proves difficult to make-well informed and secure decisions about technological trends and the life expectancy of new media." In practice, dealing with media longevity means that content needs to be copied periodically to new media or new media types. This process is called "refreshing," and is one of the activities associated with good data management practice, like the making of regular backups.

Media deterioration, however, is not the only technical preservation issue. As we noted before, a more pressing problem – and ultimately one that is more difficult to solve – is dealing with the technological obsolescence of hardware, software and media types. As Brichford and Maher (1995) pointed out with regard to hardware obsolescence, a "twenty-year life for the plastic backing material used for computer tapes and disks is irrelevant if the tape or disk drives on which they were recorded become obsolete and unavailable after ten years." The dependencies of digital objects on hardware and software can be complex. For example, Thibodeau (2002, p. 6) views digital objects as inheriting properties from three separate classes:

> Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly. A physical object is simply an inscription of signs on some physical medium. A logical object is an object that is recognised and processed by software. The conceptual object is the object as it is recognized and understood by a person, or in some cases recognized and processed by a computer application capable of executing business transactions.

Strategies for dealing with the technical obsolescence problems include the periodic migration of objects to new formats and attempts to preserve

or emulate technology. These will be introduced in more detail in the section on preservation strategies.

8.2.2 Other Challenges

In addition to these largely technical problems, there are a series of related challenges that relate to the long-term preservation of digital objects.

The first relates to the difficulties of ensuring the authenticity and integrity of objects over time. Digital information is relatively easy to manipulate, meaning that it can easily be deliberately or accidentally corrupted (Lynch 1996). The users of digital resources need to have confidence in the authenticity of preserved objects, i.e., that they are what they claim to be and that their integrity has not been compromised. There are technical methods available for dealing with this issue at the bit-level (e.g., cryptographic techniques), but confidence in an object's authenticity will ultimately be based on the level of trust a user has in the repository responsible for maintaining it.

A second problem relates to scale, i.e., the massive (and growing) amounts of digital information now being generated, combined with a proliferation of format types. The Web is but one exemplar of this, another being the "data deluge" now apparent in many scientific disciplines, whereby vast amounts of data are being generated by high-throughput instruments or streamed from sensors or satellites (Hey and Trefethen 2003; Szalay and Gray 2006). Because Web archives tend to collect multiple snapshots of Web content, they can grow very quickly indeed. For example, the largest current initiative, the Internet Archive, provides access to approximately 2 PB of data and is growing at the rate of 20 TB a month.[1] On the other hand, some national Web domain crawls can have comparatively modest storage requirements. For example, Hakala (2004) reported that a crawl of the Finnish Web domain in 2002 collected a total of 500 GB. A crawl of the Portuguese Web in 2003 processed 3.8 million URLs and downloaded 78 GB of data (Gomes and Silva 2005). By contrast, the first domain harvest of the Australian Web in 2005 took six weeks and captured 185 million documents or 6.69 TB of data (Koerbin 2005).

A final set of challenges relate to the legal contexts of digital preservation. So, for example, intellectual property rights (IPR) legislation or restrictive licensing mechanisms can sometimes restrict the collecting and preservation activities of cultural heritage organisations. Indeed, López Borrull and Oppenheim (2004) have noted that recent changes in IPR law

[1] Figures are taken from: Internet Archive, Frequently Asked Questions. Retrieved May 31, 2006 from http://www.archive.org/about/faqs.php

have tilted the balance of rights away from users in favour of content owners. While carefully constructed legal deposit legislation can help to solve some of these challenges, some preservation strategies depend on the adaptation (or reengineering) of application programs in ways that would not normally be permitted by software licenses. IPR issues are, however, not the only ones that are relevant. A detailed study of the legal contexts of Web archiving by Charlesworth (2003) noted significant problems with the potential liability of archives for providing access to defamatory or otherwise illegal content, or for breaches of data protection laws. As it is unlikely that all of these legal challenges are going to be solved in the short term, it is important that those responsible for Web archiving and digital preservation activities maintain a watching brief on legal developments in their respective legal jurisdictions.

8.3 Developing Trusted Digital Repositories

The challenge of technical obsolescence means that traditional preservation activities focused on maintaining and conserving objects are no longer effective for supporting the long-term preservation of digital information. Instead, organisations that need to preserve digital information have to develop processes and systems that can be implemented over a long period of time, adapting to external developments as necessary. This "active" approach to preservation is typified by the definition of digital preservation proposed by a working group sponsored by the Research Libraries Group (RLG) and Online Computer Library Center (OCLC): "the managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents" (RLG/OCLC Working Group on Digital Archive Attributes 2002, p. 3). These managed activities depend upon the existence of an organisational entity that can take responsibility for maintaining digital objects. In practice, this means developing some kind of repository or archive. The *Trusted Digital Repositories* report, produced by the working group referred to previously, defines a trusted repository as, "one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future" (RLG/OCLC Working Group on Digital Archive Attributes 2002, p. 5). Such repositories have to undertake a number of different functions. A start in defining some of these has been made by the OAIS Reference Model.

8.3.1 The OAIS Reference Model

The Reference Model for an Open Archival Information System (OAIS) is an attempt to provide a high-level framework for the development and comparison of digital archives or repositories (CCSDS 650.0-B-1, 2002). This standard – which has also been approved by the International Organization for Standardization as ISO 14721:2003 – was developed by the Consultative Committee for Space Data Systems (CCSDS) as part of an initiative to develop standards that would support the long-term preservation of data retrieved from satellites and other kinds of space mission. Despite these domain-specific origins, OAIS has been developed as a generic model, applicable in many other digital preservation contexts.

The OAIS model aims to provide a common framework that can be used to help understand archival challenges, especially those that relate to digital information objects. The value of the model is in providing a high-level common language that can facilitate discussion across many different communities interested in digital preservation. The standard itself defines an OAIS as an organisation of people and systems that have "accepted the responsibility to preserve information and make it available for a Designated Community" (CCSDS 650.0-B-1 2002, p. 1-11). The Working Group on Digital Archive Attributes (2002) commented that this understanding of an archive as a system of people and systems meant that the OAIS model built a stage for a better understanding of the full requirements of digital repositories.

Before exploring the functions of an OAIS in more detail, the standard defines six mandatory responsibilities that should be discharged by an archive (CCSDS 650.0-B-1 2002, p. 3-1):

- Negotiating with and accepting appropriate information from producers;
- Obtaining sufficient control of the information provided to enable its long-term preservation;
- Determining which communities constitute the "designated community" – an OAIS concept for an identified group of potential consumers (users) who should, therefore, be able to understand the information;
- Ensuring that the information to be preserved is independently understandable – i.e., without the assistance of experts – by the designated community;
- Following documented policies and procedures that ensure that information can be preserved and disseminated in an authentic way;
- Making the preserved information available to the designated community.

Much of the rest of the standard is taken up with the detailed specification of two models that detail the functional entities needed by an OAIS

and the types of information that are exchanged and managed within it. The OAIS information model will be outlined in more detail in the section on metadata below.

The functional model outlines the range of activities that would need to be undertaken by a repository, and defines in more detail those functions described within the OAIS, in order to aid the future designers of systems and to provide a set of terms and concepts for the discussion of current systems. It defines six functional entities, each of which is broken down into more detail in Unified Modeling Language (UML) diagrams.

- Ingest – accepts submissions from producers and prepares them for storage and management within the archive;
- Archival storage – for the storage, maintenance and retrieval of archive content;
- Data management – for managing information about the archive and its holdings;
- Administration – for the overall operation of the archive system;
- Preservation planning – monitoring the environment of the OAIS to ensure the long-term preservation of archive content;
- Access – supporting consumers (users) in finding and retrieving archive content.

8.3.2 Trusted Digital Repositories and Certification

The RLG/OCLC Working Group on Digital Archive Attributes (2002) built upon the foundations laid by the OAIS Model by developing a more detailed set of requirements for trusted digital repositories. They developed a set of seven attributes of trusted repositories, the first of which is compliance with the OAIS Reference Model (p. 13):

A trusted digital repository will make sure the overall repository system conforms to the OAIS Reference Model. Effective digital archiving services will rely on a shared understanding across the necessary range of stakeholders of what is to be achieved and how it will be done.

According to the OAIS standard itself, a conforming archive would fulfil the six mandatory responsibilities and support the information model it defines (CCSDS 650.0-B-1, 2002 1-3). However, it is careful to emphasise that, as a reference model, it does not define or require any particular method of implementation for either. Other attributes identified by the working group largely focus on organisational requirements that lie outside the scope of the OAIS model. They include the need for repositories to

demonstrate a fundamental commitment to apply standards and best practice (administrative responsibility), to be able to prove their organisational viability and financial sustainability, and to have a technological infrastructure appropriate for implementing suitable preservation strategies (technological and procedural suitability).

The working group also raised the question of audit and certification, recommending the development of a framework and process to support the certification of repositories.

This was the focus of a subsequent task force supported by RLG and the National Archives and Records Administration (NARA). In late 2005, this task force published a draft audit checklist for the certification of trusted digital repositories (RLG-NARA Task Force on Digital Repository Certification 2005), the use of which is currently being evaluated by the Center for Research Libraries (CRL) Audit and Certification of Digital Archives project (e.g., Dale 2005). The UK Digital Curation Centre is also collaborating with RLG by conducting audits of two repositories using the checklist. The DCC team responsible for this task argues that these audits are "designed to validate not just the appropriateness of the checklist, but to provide us with an understanding of the process and costs of its use as an audit tool" (Ross and McHugh 2005).

8.4 Digital Preservation Strategies

The OAIS Model identifies the main functions that need to be undertaken by preservation services and defines an information model for the objects held by them. However, it does not prescribe the adoption of any particular preservation strategy. This section will introduce the main range of strategies currently proposed for supporting digital preservation and comment on their appropriateness for the preservation of Web content.

The appropriateness of a given strategy depends upon the nature of the object being preserved and the reasons why it is being preserved. This means that the choice of a particular strategy, or the exact way that it is implemented, needs careful and expert consideration by repositories. In this regard, it is interesting that a number of experimental decision support tools for preservation strategies are now being developed by research projects like the digital preservation cluster of the DELOS Network of Excellence on Digital Libraries (Verdegem and Slats 2004; Rauch and Rauber 2004). Other research projects are investigating ways of dynamically implementing preservation strategies with the support of Semantic Web technologies

for the automatic detection of format obsolescence and other kinds of incompatibility (Hunter and Choudhury 2006).

Thibodeau (2002) has developed a spectrum of preservation strategies ranging on a continuum from the preservation of technology to the preservation of objects. The preserving technology side of this spectrum includes strategies based on maintaining original technologies or emulation. On the preserving objects side are approaches involving levels of abstraction like the persistent archives concept developed by a research group based at the San Diego Supercomputer Center (Moore et al. 2000). Between these two extremes are strategies based on the periodic transformation (or migration) of data. The following sections will introduce selected approaches in more detail.

8.4.1 Preserving Technology

On the face of it, the simplest preservation strategy would be to keep and maintain all original application programs, operating systems and hardware platforms. Lee et al. (2002) comment that advocates of this strategy argue that it is the only way of preserving the *behaviour* as well as the look and feel of a given digital object. However, this approach crumbles in the face of rapid technology obsolescence and the impossibility of maintaining hardware over long periods of time. Feeny (1999, p. 42) has argued that this strategy would quickly result in the existence of museums of "ageing and incompatible computer hardware." While the strategy may have some value where the hardware is particularly unique or historically significant, this is certainly not an approach that is appropriate for Web content, which is usually accessed via browser software rather than being dependent on any specific type of hardware.

8.4.2 Emulating Technology

Preservation strategies based on emulating technology abandon attempts to keep obsolete hardware working and focus instead on the development of programs that enable the continued use of application programs in new environments. A basic assumption of the strategy is that digital resources are inherently software-dependent. According to Rothenberg (1999, p. 8), "digital documents exist only by virtue of software that understands how to access and display them; they come into existence only by virtue of running this software." He argues that the only reliable way of recreating a document's original functionality, look, and feel would, therefore, be "to enable the emulation of obsolete systems on future, unknown systems" (p. 17). The

importance of the emulation approach is that it enables the preservation of digital objects in their original forms, which aids their authenticity.

Emulation, therefore, is based on the development of software programmes (emulators) that mimic the behaviour of obsolete hardware. Technically speaking, this is far from being a trivial task, but the fact that hardware tends to be well specified at a logical level means that it is an easier task than reengineering application software for new computing environments (Digital Preservation Testbed 2003). Once an emulation approach has been chosen, there is a need to solve the practical question of exactly how emulator programs will be run on future generations of hardware. One suggested approach is to "rehost" emulator programs periodically onto new hardware platforms, which could be quite resource intensive. An alternative is "chaining," whereby emulator platforms are in time themselves successively emulated, enabling previous emulators to be run under a chain of emulators. Another approach that has been suggested involves developing all emulator programs to function on a virtual platform – an "emulation virtual machine" – that can in turn be successively implemented on new hardware platforms (Rothenberg 2000). The advantage of this approach is that it simplifies the amount of rehosting required, as only the virtual machine itself needs to be rewritten. A variant of the virtual machine approach is the Universal Virtual Computer (UVC) concept developed by Raymond A. Lorie of IBM (Lorie 2002). A UVC is a simple general-purpose computer that can be implemented as a platform independent layer on current and future hardware. Rather than running original application programs, formats supported by the UVC are decoded into a Logical Data Schema that can be used with format decoders in the future to reconstruct objects on future implementations of the UVC. The National Library of the Netherlands and IBM has tested the concept through the development of a UVC demonstration tool for JPEG images (Hoeven et al. 2005).

Emulation would, in principle, seem to represent an appropriate preservation approach for at least some Web content. Web browsers tend to be implemented on a wide range of different platforms, so the exact choice of hardware to emulate would to some extent be arbitrary. Because Web standards and the technologies supported by browsers change over time, it may be important to maintain multiple versions of browser software in order to ensure that pages can be rendered in an appropriate manner. It will also be important to maintain copies of browser "plug-ins" and related technologies.

8.4.3 Migrating Objects

The data migration preservation strategy abandons any attempt to keep obsolete hardware or application programs working, even in surrogate form. Instead, data objects are continually transferred in order to work on new generations of hardware and software. The Task Force on Archiving of Digital Information (Garrett and Waters 1996, p. 6) provides a much-cited definition:

> Migration is the periodic transfer of digital materials from one hardware/-software configuration to another, of from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.

Data migration is currently the most tried-and-tested preservation approach, often combined with some kind of format standardisation undertaken on ingest. Migration has been used for decades by the computer industry to ensure that current information remains accessible and usable. It also underlies the versioning behaviour of formats in many office programs. From a longer-term perspective, however, it suffers from a number of problems. First, because objects are subject to almost continuous change, it is very difficult to ensure that they retain their authenticity (Tibbo 2003, p. 22). Second, migration processes are not particularly efficient for large collections of heterogeneous objects, which would need constant monitoring and intervention. Rothenberg (1999, p. 13) argues that migration approaches are labour-intensive, "time-consuming, expensive, error prone, and fraught with the danger of losing or corrupting information."

To help overcome some of these problems, other variants of the migration approach have been developed. For example, researchers based at the University of Leeds (Mellor et al. 2002) proposed a form of "migration on demand," whereby an object's original bit-stream would be preserved – helping to maintain its authenticity – and migrated only at the point of delivery. In this model, the focus of migration moves on to the migration tools, rather than the objects themselves.

In principle, data migration approaches could be applied to the majority of Web content. The experiences of Web archiving initiatives to date suggests that much of the surface Web – at least – is made up of a relatively limited number of formats. For example, a crawl of the Finnish Web in June 2002 found that over 96% of harvested content was in one of four

formats: HTML (48%), GIF (25%), JPEG (20%) and PDF (3%).[2] While it would be possible to develop data migration strategies for all of these formats – and any others that become important – the real problem is that the user experience of the Web as a whole is not so easy to migrate. The exact role of migration in supporting the preservation of Web content remains, therefore, an open question.

8.4.4 Other Strategies

A number of other preservation strategies exist. Some of these are based on the concept of encapsulation, the idea that preserved objects should essentially be self-describing, linking content with all of the information required for it to be deciphered and understood. As we will see, this is the heart of the idea that underlies the Information Package concept in the OAIS reference model. It also underlies initiatives like the Universal Preservation Format (Shepard 1998) and the self-documenting encapsulation concept developed as part of the Victorian Electronic Records Strategy (Waugh et al. 2000; Waugh 2006).

Another type of approach is exemplified by the persistent-archives concept developed by the San Diego Supercomputing Centre as part of a series of research projects funded by NARA and other agencies (Moore et al. 2000). This is based on the development of a complete preservation infrastructure that enables the preservation of the organisation of collection as well as the objects that make up that collection, maintained in platform-independent form. The architecture used enables any hardware or software component to be replaced with minimal effect on the rest of the system.

8.4.5 Final Thoughts on Preservation Strategies

The existence of multiple approaches reflects the reality that we do not really know yet which strategies will work best for a given object or preservation objective. They are also not mutually exclusive, meaning that risk can be spread across a number of different strategies. The key thing, whichever strategy (or combination of strategies) is chosen is to understand that the purpose of any strategy will be to ensure that the significant properties of preserved objects can be retained. In the short to medium term, however, much more research into preservation strategies will be needed (e.g., Hedstrom 2002; Tibbo 2003; Ross and Hedstrom 2005).

[2] These figures are taken from a survey undertaken as part of a feasibility study into Web archiving in the UK (Day, 2003)

8.5 Preservation Metadata

The key to the successful implementation of all preservation strategies will be the capture, creation, maintenance and application of appropriate metadata (e.g., Day 2004; 2005). This "preservation metadata" is understood to be all of the *various types of data* that allows the re-creation and interpretation of the structure and content of digital data over time (Ludäscher et al. 2001). Understood in this way, it is clear that such metadata needs to support an extremely wide range of different functions, including discovery and access, recording the contexts and provenance of objects, to the documentation of repository actions and policies. Conceptually, therefore, preservation metadata spans the traditional division of metadata into descriptive, structural and administrative categories. Lynch (1999), for example, has noted that within digital repositories, metadata should accompany and make reference to digital objects, providing associated descriptive, structural, administrative, rights management, and other kinds of information.

The wide range of functions that preservation metadata is expected to support means that the definition (or recommendation) of standards is not a simple task. The situation is complicated further by the knowledge that different kinds of metadata will be required to support different digital preservation strategies and that the metadata standards themselves will need to evolve over time.

To date, the information model defined by the OAIS Reference Model has been extremely influential on the development of preservation metadata standards.

8.5.1 The OAIS Information Model

The OAIS information model defines two main categories of metadata that needs to be associated with the objects that need information. First all Information Objects handled by an archive (content, metadata, etc.) are made up of a Data Object – which for digital objects would typically be a sequence of bits – and the associated Representation Information needed to permit the full conversion of these bits into meaningful information (CCSDS 650.0-B-1 2002, p. 4-19). The OAIS model defines this Representation Information as "the information that maps a Data Object into more meaningful concepts" (CCSDS 650.0-B-1 2002, p. 1-13), but for digital resources it is essentially the technical information (or metadata) needed to render the bit sequences into something that can be read or used by its designated community. Typically, Representation Information might

include descriptions of the formats, character sets, etc. in use, possibly including software and descriptions of hardware and software environments. In the OAIS model, this is known as Structure Information. It might also include any additional information that is required to establish the particular meaning of data content, e.g., that raw numbers should be understood as dates or as temperatures in degrees Celsius. The OAIS model refers to this as Semantic Information. The OAIS information model understands that Representation Information can be recursive, i.e., that it may itself may need Reference Information in order to be made meaningful, resulting in what the model calls as a Representation Network. While Representation Information is conceptually part of the Content Information, in practice its presence could be fulfilled by a link to centralized information held elsewhere within the OAIS or even in third party registries. A start has been made with developing registries of information about file formats (Abrams and Seaman 2003; Darlington 2003), but similar approaches could be used for other types of Representation Information (Giaretta et al. 2005).

The model also encapsulates Content Information with additional metadata – known as Preservation Description Information (PDI) to form an entity known as an Information Package. The standard defines several types of Information Package (e.g., for submission into the archive and for dissemination), but the most significant for preservation purposes is the Archival Information Package (AIP), "defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object" (CCSDS 650.0-B-1 2002, p. 4-33). In short, the archival information package is a way of conceptually linking the object that is the primary focus of preservation together with *all* of the additional types of information (or metadata) that are necessary to support its continued use over time.

The OAIS information model says that Preservation Description Information is "specifically focused on describing the past and present states of the Content Information, ensuring that it is uniquely identifiable, and ensuring that it has not been unknowingly altered" (CCSDS 650.0-B-1 2002, p. 4-27). The Information Model defines four separate classes of PDI, broadly based on categories defined in the report of the Task Force on Archiving of Digital Information, namely: fixity, reference, context and provenance. The report (Garrett and Waters 1996) noted that these four categories, together with the definition of content at different levels of abstraction, were the key features for determining information integrity in the digital environment and argued that they deserved special attention.

- *Fixity* – We have already mentioned that the users of digital resources need to have confidence that they are what they claim to be and that their integrity has not been compromised. While metadata by itself cannot solve all integrity problems, the OAIS model suggests the inclusion of Fixity Information that can support data integrity checks at the level of Content Data Objects. These might include the use of cryptographic techniques like checksums that can help protect bit-level integrity by highlighting any changes made to individual data objects.
- *Reference* – Another aspect of integrity identified by the Task Force on Archiving of Digital Information was the need for objects to be identified and located over time. Their report said that for an object "to maintain its integrity, its wholeness and singularity, one must be able to locate it definitively and reliably over time among other objects" (Garrett and Waters 1996, p. 15). This brings us to the traditional realm of descriptive metadata, e.g., that used in bibliographies, catalogues, and finding aids, but also highlights a key role for persistent identifiers. Identifiers feature highly in the OAIS model's definition of Reference Information, although the practical examples make it clear that other types of descriptive metadata could also be included. There is a separate category in the OAIS information model for descriptive metadata about information packages (Descriptive Information) that can be used to facilitate discovery and access, although it acknowledges that at least some Reference Information will often be replicated in these Package Descriptions.
- *Context* – Many resources cannot properly be interpreted without some understanding of their context. Digital objects do not often exist in isolation, but interact with other objects and their wider environment. This is especially true of Web resources. The context might, for example, be technical, e.g., recording dependencies on particular hardware or software configurations. In the OAIS information model, Context Information is defined as documenting the relationships of the Content Information to its environment (CCSDS 650.0-B-1 2002, p. 4-28).
- *Provenance* – The OAIS model understands Provenance Information as a specific type of Context Information that documents the history of the Content Information. This might include information about its creation and provide a record of custody and preservation actions undertaken. Provenance also refers to a longstanding principle of the archives profession and embodies the concept that a key part of the integrity of an object is being able to trace its origin and chain of custody (Tibbo 2003, p. 32).

8.5.2 The PREMIS Data Dictionary and Other Standards

The first metadata specifications specifically designed to address digital preservation requirements were developed in the late 1990s by the National Library of Australia and European research projects like Cedars (CURL Exemplars in Digital Archives) and NEDLIB (Networked European Deposit Library) (e.g., Day 2001). Between 2000 and 2002, an international working group commissioned by OCLC and RLG built upon these (and other) proposals to produce a unified *Metadata Framework to Support the Preservation of Digital Objects* (OCLC/RLG Working Group on Preservation Metadata 2002). This Metadata Framework was *explicitly* structured around the OAIS information model, defining various metadata elements for Content Information (including Representation Information) and PDI.

Following publication of the Metadata Framework, OCLC and RLG commissioned a further working group to investigate the issues of implementing preservation metadata in more detail. The resulting Working Group on Preservation Metadata: Implementation Strategies (PREMIS) had the twin objectives of producing a "core" set of preservation metadata elements and evaluating alternative strategies for encoding, storing, managing and exchanging such metadata.

The working group issued its proposal for core preservation metadata elements in May 2005 with the publication of the PREMIS *Data Dictionary for Preservation Metadata* (PREMIS Working Group 2005). While this is intended to be a translation of the earlier Metadata Framework into a set of implementable semantic units, the Data Dictionary developed its own data model and was not afraid to diverge from the OAIS model in its use of terminology. The Data Dictionary defines preservation metadata as "the information a repository uses to support the digital preservation process," specifically that "metadata supporting the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" (PREMIS Working Group, 2005, p. ix). The Data Dictionary itself defines elements (semantic units) for describing four of the entities identified by the PREMIS data model: objects (at different levels of aggregation), events, agents and rights, the latter two in no real detail. The working group also limited the scope of the Data Dictionary by excluding categories of metadata deemed not directly relevant to preservation (e.g., descriptive metadata) or outside the expertise of the group (e.g., technical information about media and hardware).

The examples included in the Data Dictionary include a snapshot of a website. This gives an indication of the potential complexity of PREMIS, in particular when it is applied to Web content.

8.5.3 Web Archiving and Metadata

Some Web archiving initiatives already collected some simple metadata. For example, initiatives using crawler programs for domain capture of the surface Web record certain aspects of documents harvested. These might include a document's original URL, its checksum, and a record of the time the document was harvested. Hakala (2004) describes how these can be used for duplicate detection and other Web archive management processes. The IIPC is currently working on the development of a Web Archiving Metadata Set that would define the richer metadata that can be automatically generated or captured by IIPC tools, e.g., metadata about harvesting parameters, website contexts, etc.

8.6 Digital Preservation and the Web

Before concluding, it might be worth outlining briefly some of the reasons why the Web may prove to be a particularly difficult object to preserve.

First, the Web is a deceptively complex object. In governance terms it remains what Strogatz (2004, p. 255) calls an "unregulated, unruly labyrinth where anyone can post a document and link it to any page at will." The result of this is hidden complexity. For example, the surface Web alone links a wide range of document types (e.g., text, images, sound, multimedia, software) in an even wider range of formats – all of which may need to be considered separately from a preservation perspective. The Web also includes (or provides interfaces to) databases, digital libraries, metadata collections, and interactive sites like "Web logs." In addition, while some website behaviour is determined at the server side (Fitch 2003), other aspects of functionality depend on the exact combination of browser software and "plug-ins" available to the user. In this context, it is difficult for preservation initiatives to make decisions about the significant properties and authenticity of objects. Lyman (2002, p. 41) argues that, for authenticity, preserved documents "must both include the context and evoke the experience of the original."

A related problem is the Web's dynamic nature. Web archiving initiatives can only preserve "snapshots" of sites or domains at the expense of their dynamism, rather like insects trapped in amber. Once snapshots of Web content are located outside the active Web, it is arguably missing one of its most characteristic properties (Tibbo 2003, p. 16).

The problems of complexity and dynamism reflect a deeper lack of clarity on defining the exact boundaries of the Web. It is a general principle of digital preservation that it is important to understand exactly what is being

preserved in order to preserve it most effectively. On detailed examination, however, the Web can be a fairly nebulous concept. For example, many "hidden Web" sites just provide browser-friendly access to a managed database whose content often predates the Web, and will most likely survive it (see Chap. 5). We may need to ask whether these particular sites fall within the scope of Web archiving initiatives as currently constituted, or whether they should be dealt with in other ways.

8.7 Conclusion

This chapter has attempted to introduce some of the range of managed activities that are necessary to ensure the long-term preservation of collections of Web content. It has focused in particular on the development of trusted repository systems and the adoption of appropriate digital preservation strategies, noting the key role of metadata. Digital preservation has been described as a grand challenge for the first decade of the twenty first century (Tibbo 2003). Preserving Web content for the long-term promises to be one of the most demanding parts of this challenge.

Acknowledgements

UKOLN is funded by the Joint Information Systems Committee (JISC) and the Museums, Libraries and Archives Council (MLA), as well as by project funding from various sources. UKOLN also receives support from the University of Bath, where it is based.

Parts of this paper are based on a presentation given at the seminar *Archivo de la Internet española: Webs y archivos personales*, held at the Residencia de Estudiantes, Madrid on December 12, 2005.

References

Abrams, S. L. & Seaman, D. (2003). *Towards a global digital format registry*. Paper presented at the 69th IFLA General Conference and Council, Berlin, Germany, August 1–9, 2003. Retrieved May 31, 2006 from http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Bar-Ilan, J. & Peritz, B. C. (2004). Evolution, continuity, and disappearance of documents on a specific topic on the web: a longitudinal study of 'informetrics'. *Journal of the American Society for Information Science and Technology*, *55*(11), 980–990

Brichford, M. & Maher, W. (1995). Archival issues in network electronic publications. *Library Trends*, *43*(4), 701–712

Brygfjeld, S. A. (2002). Access to Web archives: the Nordic Web Archive Access Project. *Zeitschrift für Bibliothekswesen und Bibliographie*, *49*, 227–231

CCSDS 650.0-B-1. (2002). Reference model for an Open Archival Information System (OAIS). Retrieved May 31, 2006 from Consultative Committee on Space Data Systems Web site: http://public.ccsds.org/publications/archive/650x0b1.pdf

Charlesworth, A. (2003). *A study of legal issues related to the preservation of Internet resources in the UK, EU, USA and Australia*. Retrieved May 31, 2006 from Joint Information Systems Committee Web site: http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf

Crichlow, R., Davies, S., & Wimbush, N. (2004). Accessibility and accuracy of Web page references in 5 major medical journals. *JAMA: the Journal of the American Medical Association*, *292*(22), 2723–2724

Dale, R. L. (2005). Making certification real: developing methodology for evaluating repository trustworthiness. *RLG DigiNews*, 9(5). Retrieved May 31, 2006 from http://www.rlg.org/en/page.php?Page_ID=20793

Darlington, J. (2003). PRONOM – a practical online compendium of file formats. *RLG DigiNews*, *7*(5). Retrieved May 31, 2006 from http://www.rlg.org/preserv/diginews/diginews7-5.html

Day, M. (2001). Metadata for digital preservation: a review of recent developments. In P. Constantopoulos & I. Sølvberg (Eds.), *Research and advanced technology for digital libraries, 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4–9, 2001* (pp. 161–172). Lecture Notes in Computer Science, 2163. Berlin Heidelberg New York: Springer

Day, M. (2003). *Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Welcome Trust*. Retrieved May 31, 2006 from Joint Information Systems Committee Web site: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

Day, M. (2004). Preservation metadata. In G. E. Gorman & D. G. Dorner (Eds.), *Metadata applications and management* (pp. 253–273). International Yearbook of Library and Information Management, 2003-2004. London: Facet

Day, M. (2005). Metadata. In S. Ross & M. Day (Eds.), *DCC Digital Curation Manual*. Retrieved May 31, 2006 from Digital Curation Centre Web site: http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/

Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M., & Schilling, L. M. (2003). Going, going, gone: lost Internet references. *Science*, *302*, 787–788

Digital Preservation Testbed. (2003). *Emulation: context and current status*. Retrieved May 31, 2006 from Nationaal Archief Web site: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf

Feeny, M. (1999). *Digital culture: maximising the nation's investment*. London: National Preservation Office

Fitch, K. (2003). *Web site archiving: an approach to recording every materially different response produced by a Website*. Paper presented at the 9th Austral-

asian World Wide Web Conference, AusWeb03, *Sanctuary Cove, Queensland, Australia, July 5–9, 2003*. Retrieved May 31, 2006 from http://ausweb.scu.edu.au/aw03/papers/fitch/

Garrett, J. & Waters, D. (1996). *Preserving digital information: report of the Task Force on Archiving of Digital Information*. Washington, DC: Commission on Preservation and Access; Mountain View, CA: Research Libraries Group. Retrieved May 31, 2006 from http://www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf

Giaretta, D., Rankin, S., McIlwrath, B., Rusbridge, A. & Patel, M. (2005) Representation Information for interoperability now and with the future. In *Local to global data interoperability – challenges and technologies, IEEE Mass Storage Systems & Technology Committee, Sardinia, Italy, June 20–24, 2005* (pp. 42–46). Piscataway, NJ: Institute of Electrical and Electronics Engineers

Gomes, D. & Silva, M. J. (2005). Characterising a national community Web. *ACM Transactions on Internet Technology*, *5*(3), 508–531

Hakala, J. (2004). Archiving the Web: European experiences. *Program*, *38*(3), 176–183

Hedstrom, M. (1998). Digital preservation: a time bomb for digital libraries. *Computers and the Humanities*, *31*(3), 189–202

Hedstrom, M. (2002). The digital preservation research agenda. In *The state of digital preservation: an international perspective* (pp. 32–37). Washington, DC: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub107abst.html

Hester, E. J., Heilig, L. F., Drake, A. L., Johnson, K. R., Vu, C. T., Schilling, L. M., & Dellavalle, R. P. (2004). Internet citations in oncology journals: a vanishing resource? *Journal of the National Cancer Institute*, *96*(12), 969–971

Hey, T. & Trefethen, A. (2003). The data deluge: an e-science perspective. In F. Berman, G. Fox & A. J. G. Hey (Eds.), *Grid computing: making the global infrastructure a reality* (pp. 809–824). Chichester: Wiley

Hoeven, J. R. van der, Diessen, R. J. van, & Meer, K. van der. (2005). Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects. *Journal of Information Science*, *31*(3), 196–208

Hunter, J. & Choudhury, S. (2006). PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries*, *6*(2), 174–183

ISO 14721:2003: Space data and information transfer systems – Open archival information system – Reference model. Geneva: International Organization for Standardization

Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research*, 9(2), 174. Retrieved May 31, 2006 from http://informationr.net/ir/9-2/paper174.html

Koerbin, P. (2005). *Report on the crawl and harvest of the whole Australian Web domain undertaken during June and July 2005*. Retrieved May 31, 2006 from National Library of Australia Web site: http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf

Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. Å., Kruger, A., & Giles, C. L. (2001). Persistence of Web references in scientific research. *Computer*, *34*(2), 26–31

Lee, K.-H., Slattery, O., Lu, R., Tang, X. & McCrary, V. (2002). The state of the art and practice in digital preservation. *Journal of Research of the National Institute of Standards and Technology*, *107*, 93–106

López Borrull, A. & Oppenheim, C. (2004). Legal aspects of the Web. *Annual Review of Information Science and Technology*, *38*, 483–548

Lorie, R. A. (2002). *The UVC: a method for preserving digital documents.* Amsterdam: IBM Netherlands. Retrieved May 31, 2006 from Koninklijke Bibliotheek Web site: http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf

Ludäscher, B., Marciano, R., & Moore, R. (2001). Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *SIGMOD Record*, *30*(3), 54–63

Lyman, P. (2002). Archiving the World Wide Web. In *Building a national strategy for digital preservation* (pp. 38–51). Washington, DC: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub106abst.html

Lynch, C. (1996). Integrity issues in electronic publishing. In R. P. Peek & G. B. Newby (Eds.), *Scholarly publishing: the electronic frontier* (pp. 133–145). Cambridge, MA: MIT

Lynch, C. (1999). Canonicalisation: a fundamental tool to facilitate preservation and management of digital information. *D-Lib Magazine*, 5(9). Retrieved May 31, 2006 from http://www.dlib.org/dlib/september99/09lynch.html

Mellor, P., Wheatley, P., & Sergeant, D. (2002). Migration on request: a practical technique for digital preservation. In M. Agosti & C. Thanos (Eds.), *Research and advanced technology for digital libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16–18, 2002* (pp. 516–526). Lecture Notes in Computer Science, 2458. Berlin Heidelberg New York: Springer

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., & Gupta, A. (2000). Collection-based persistent digital archives – part 1. *D-Lib Magazine*, 6(3). Retrieved May 31, 2006 from http://www.dlib.org/dlib/march00/moore/03moore-pt1.html

OCLC/RLG Working Group on Preservation Metadata. (2002). *A metadata framework to support the preservation of digital objects*. Dublin, Ohio: OCLC Online Computer Library Center. Retrieved May 31, 2006 from http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

PREMIS Working Group. (2005). *Data dictionary for preservation metadata* Dublin, Ohio: OCLC Online Computer Library Center. Retrieved May 31, 2006 from http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Rauch, C. & Rauber, A. (2004). Preserving digital media: towards a preservation solution evaluation metric. In Z. Chen, H. Chen, Q. Miao, Y. Fu, E. A. Fox & E. -P. Lim (Eds.), *Digital libraries: international collaboration and cross-fertilization, 7th International Conference on Asian Digital Libraries, ICADL 2004, Shanghai, China, December 13-17, 2004* (pp. 203–212). Lecture Notes in Computer Science, 3334. Berlin Heidelberg New York: Springer

RLG/OCLC Working Group on Digital Archive Attributes. (2002). *Trusted digital repositories: attributes and responsibilities*. Mountain View, CA: Research Libraries Group. Retrieved May 31, 2006 from http://www.rlg.org/legacy/longterm/repositories.pdf

RLG-NARA Task Force on Digital Repository Certification. (2005). *An audit checklist for the certification of trusted digital repositories: draft for public comment*. Mountain View, CA: RLG. Retrieved May 31, 2006 from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Ross, S. & Gow, A. (1999). *Digital archaeology: rescuing neglected and damaged data resources*. London: South Bank University, Library Information Technology Centre

Ross, S. & Hedstrom, M. (2005). Preservation research and sustainable digital libraries. *International Journal on Digital Libraries*, *5*(4), 317–324

Ross, S. & McHugh, A. (2005). Audit and certification of digital repositories: creating a mandate for the Digital Curation Centre (DCC). *RLG DigiNews*, *9*(5). Retrieved May 31, 2006 from http://www.rlg.org/en/page.php?Page_ID=20793

Rothenberg, J. (1999). *Avoiding technological quicksand: finding a viable technical foundation for digital preservation*. Washington, DC: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub77.html

Rothenberg, J. (2000). *An experiment in using emulation to preserve digital publications*. Den Haag: Koninklijke Bibliotheek. Retrieved May 31, 2006 from http://nedlib.kb.nl/results/emulationpreservationreport.pdf

Sellitto, C. (2005). The impact of impermanent Web-located citations: a study of 123 scholarly conference publications. *Journal of the American Society of Information Science and Technology*, *56*(7), 695–703

Shepard, T. (1998). Universal Preservation Format (UPF): conceptual framework. *RLG DigiNews*, 2(6). Retrieved May 31, 2006 from http://www.rlg.org/preserv/diginews/diginews2-6.html

Smith, A. (2003). *New-model scholarship: how will it survive?* Washington, D.C.: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub114abst.html

Spinellis, D. (2003). The decay and failure of Web references. *Communications of the ACM*, *46*(1), 71–77

Strogatz, S. (2004). *Sync: the emerging science of spontaneous order*. London: Penguin

Szalay, A. & Gray, J. (2006). Science in an exponential world. *Nature*, *440*, 413–414

Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. In *The state of digital preservation: an international perspective* (pp. 4–31). Washington, DC: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub107abst.html

Tibbo, H. R. (2003). On the nature and importance of archiving in the digital age. *Advances in Computers*, *57*, 1–67

Van Bogart, J. W. C. (1995). *Magnetic tape storage and handling: a guide for libraries and archives*. Washington, DC: Commission on Preservation and Access; St. Paul, Minn.: National Media Laboratory. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub54.html

Verdegem, R. & Slats, J. (2004). Practical experiences of the Dutch digital preservation test-bed. *VINE: the Journal of Information and Knowledge Management Systems*, *34*(2), 56–65

Waugh, A. (2006). The design of the VERS encapsulated object experience with an archival information package. *International Journal on Digital Libraries*, *6*(2), 184–191

Waugh, A., Wilkinson, R., Hills, B., & Dell'oro, J. (2000). Preserving digital information forever. In *ACM 2000 Digital Libraries, 5th ACM Conference on Digital Libraries, San Antonio, TX, USA, June 2–7, 2000* (pp. 175–184). New York: Association for Computing Machinery

Wren, J. D. (2004). 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*, *20*(5), 668–672

9 Year-by-Year: From an Archive of the Internet to an Archive on the Internet

Michele Kimpton and Jeff Ubois

Internet Archive
michele@archive.org
jeff@archive.org

9.1 Introduction

Since its beginnings in 1995, the Internet Archive has pursued the long-term goal of providing universal access to all knowledge, within our lifetime.

Over the last 10 years, the Archive has engaged in projects in many different countries, for many different types of users, and worked to preserve Web pages, books, music, software, and moving images, and to make them accessible via the Internet.

Today, the Archive's Wayback Machine serves 70,000 unique visitors per day and 200 requests per second. Its 600 TB collection includes 50 billion Web pages, 30,000 books, 36,0000 sound recordings, 15,000 movies.

Hundreds of individual contributors have made this possible. So have technical advances in the computer industry, where system performance doubles every 12–18 months. In the archival community, 10 years is not a long time. But in the computer industry, 10 years is enough time to create a 100-fold factor of improvement: since 1997, the price of raw disk storage has dropped by more than 99%, from $180 to under 50 cents per gigabyte, and more than 25 million broadband connections have been added in the US alone.

It is by understanding the rate of improvement in computer based storage and access that it is possible to begin to credit what Raj Reddy of Carnegie Mellon University has described as "universal access to all knowledge."

9.2 Background: Early Internet Publishing

Even by the mid 1980s, it was clear that a change was coming to the world of electronic publishing. In the first half of the 1990s, as newspapers began to move from closed, proprietary databases to the Internet, the idea of the Internet as a library began to take shape. Internet publishing systems such as Wide Area Information Servers (WAIS) and Gopher were seen as complements to Web pages; the metaphor of the Internet as a book, with Web *pages*, referenced by Gopher servers that provided a *Table of Contents*, and an *index* produced by WAIS servers was offered as an alternative to the metaphor of the Internet as a "superhighway."

The launch of the AltaVista service in December, 1995 proved that all of the pages on the Web could be treated as a single collection, and indexed and made searchable for all users on the Net. What was not clear is how records of those pages would be maintained.

9.3 1996: Launch of the Internet Archive

The Internet Archive was formally incorporated in April, 1996, by Bruce Gilliat and Brewster Kahle.

By that time, broken links (404 errors) were a growing problem, and it was clear that most Web pages were short-lived. Some solution to this problem was needed, and a system for archiving Web pages before they vanished seemed like an obvious approach.

This led to an early design decision at the Archive about the collections policy: to be aggressive about collecting material that was in danger of disappearing, and to be opportunistic about collecting and digitizing items from the past, such as Usenet postings.

Still, as of 1996, there was little sense in the Internet community that loss of Web pages was a particularly important problem. Since the Web did not have much history, it was difficult to describe uses for expired pages.

To demonstrate the potential value of such pages, the Archive partnered with the Smithsonian Institution in Washington, DC to collect snapshots of the websites of all the 1996 Presidential candidates.[1] The tools for this project were not terribly sophisticated; they were essentially PC applications built to capture entire websites by following the links from the main page.

[1] See http://movie0.archive.org/96_Elections/index.htm

This data was eventually incorporated into the Smithsonian's presidential archive, which now includes pages from five political parties, and numerous candidates for president, ranging from Bill Clinton to Pat Buchanan. Many of the sites in this archive were shut down as candidates dropped out of the race.

Based on this success, the Library of Congress commissioned the Archive to create a focused online collection of the 2000 election, and renewed this request for the 2002 elections.

Also, in 1996, the Archive began its relationship with Alexa Internet, a for-profit company that began crawling and archiving the Web in November to provide data for a browser toolbar (plug-in) offering data about sites being viewed, and based on data gathered from other users, suggestions about related pages that might also be of interest. The Internet Archive still relies on Alexa Internet to provide data from crawls of the Web.

Two other developments from 1996 are worth noting.

The first is technical. In 1996, tape still offered a considerable price advantage over disk, and the Archive built its first generation infrastructure using tape storage robots, beginning with an ADIC 50. Despite generous contributions from leading vendors, this was ultimately to prove untenable; the access requirements posed by the Archive's users were simply too intense, and retrieval times were too slow. As Bruce Gilliat humorously says: getting a page could be done in few seconds… or days later.

The second is legal. The legal implications of collecting Web pages aggressively and to serve them up on an "opt-out" basis, as the Archive began to do that year, were unclear. AltaVista's promotion of the robots.txt exclusion protocol was an important step because it shifted the "defaults" – Web page owners that wanted to avoid landing in a search engine index, or in the Archive had an easy way to opt out, or to remove pages they clearly owned (as proven by their ability to modify the root directory of the Web site in question). While robots.txt provided resolution of the removals issue for those who owned pages in the Archive, it did not resolve the question of removing pages owned by others. This removals issue is described later with reference to small conference at UC Berkeley that was initiated by the Internet Archive in 2002 (Ubois 2002).

9.4 1997: Link Structure and Tape Robots

Collection of Web pages, link data, and "usage trails," i.e., the choices made by millions of Web users as they moved from page to page, began in

earnest in 1997 with the release of Alexa's toolbar, a browser plug-in that helped users navigate the Web by providing information on the site being viewed, including suggestions about related sites.

The link data and usage trails gathered by Alexa functioned as a collaborative filtering system, highlighting the pages that the Internet community as a whole valued most highly. Links and clicks were essentially votes on the value of a given page.

The ability to automatically determine page value was closely related to another critical design decision.

Some of the largest and most successful Web "libraries" in 1997 were essentially catalogs of sites like Yahoo. But it was uncertain how scalable a manual cataloging approach could be over time. Might it be possible to eliminate manual cataloging and the entire selections process in favor of a "collect it all" approach, combined with user-generated metadata in place of a catalog?

The answer seemed to be yes, and the Alexa began crawling pages according to how much usage they received, as measured by data gathered from its browser toolbar. Pages visited most frequently by people were backed up first.

The Alexa crawler was set to make a snapshot of the Web every eight weeks, and that schedule is still in place, though the size of each crawl increased to 100 TB in 2004 from 1 TB in 1997.

The other issue facing the Archive in 1997 was whether to rely on tape or disk storage. Tape still had cost advantages over disk, but access was slow.

As noted in (Gray and Shenoy 1999) "tape, disk, and RAM have maintained price ratios of about 1:10:1000. That is, disk storage has been 10x more expensive than tape, and RAM has been 100× more expensive than disk."

But when access costs are used as the yard stick, disk as actually much cheaper: "a tape archive is half the cost per terabyte of disk storage, but tape does not provide easy access to the data. The cost per random tape access is about a hundred thousand times higher (100 accsses/s/1 K$ disk vs. 000.01 accesses/second/10,000$ tape) (Gray and Shenoy 1999).

9.5 1998: Getting Archive Data Onto (Almost) Every Desktop

Between 1996, when Netscape went public, and 1998, the Internet and Internet-related businesses became the focus of a worldwide shift in investment priorities. As billions of dollars poured into public markets,

venture funds, and Internet startups, the number of pages to archive was doubling every 3–6 months. The number of users on the Internet was also doubling every few months. Access was becoming a concern.

In an effort to provide access to its holdings, and to establish the Archive and Alexa Internet as part of the Internet's infrastructure, Alexa entered into contracts with Microsoft and Netscape to bundle its software into the Internet Explorer and Netscape browsers. This gave Alexa presence on 90% of the world's desktop computers, and whether users knew or not, it gave them access to data held by the Internet Archive.

For the Archive, the need to begin serving up data for tens of millions of users put a strain on the tape-based infrastructure.

By the end of 1998, two things were clear:

- Scaling up would require a move from tape to disk. As the number of access requests increased, the ability of tape robots to respond was proving to be inadequate;
- Manual collections policies were more expensive than disk space; that is, archiving based on hand cataloging particular websites was more expensive than simply archiving all accessible sites according to data gathered from end users by Alexa.

9.6 1999: From Tape to Disk, A New Crawler, and Moving Images

The commercial success of the Alexa service, determined in part by its presence on virtually every PC connected to the Internet, led Amazon to acquire Alexa in 1999. This ultimately led to changes in the structure of both organizations.

An important technical development in 1999 was the creation of a new crawler by Andy Jewel. The new crawler was better able to handle parallel processes for gathering Web data, and was manageable across multiple machines. This crawler enabled Alexa to filter out 16 billions URL to crawl 4 billions and to expand the breadth and depth of its crawl.

It also cemented the decisions around the ARC file format used to store Web pages. Originally developed in 1996 by Mike Burner and Brewster Kahle, the ARC File Format Specification (Burner and Kahle 1996), which was designed to meet several requirements:

- The file must be self-contained: it must permit the aggregated objects to be identified and unpacked without the use of a companion index file;

- The format must be extensible to accommodate files retrieved via a variety of network protocols, including HTTP, FTP, news, gopher, and mail;
- The file must be "streamable": it must be possible to concatenate multiple archive files in a data stream;
- Once written, a record must be viable: the integrity of the file must not depend on subsequent creation of an in-file index of the contents.

While the specification does not require an external index of the contents and object-offsets, such an index greatly enhances the retrievability of objects stored in this format. Today, the Archive maintains such indices, and is seeking to standardize their format through the International Internet Preservation Consortium.[2]

1999 also marked a move beyond Web pages into other types of data. By 1999, storage prices had dropped to the point at which the Archive could begin collecting moving images. Through a partnership with Rick Prelinger of the Prelinger Archives, a project to digitize 1,000 films (for an ultimate cost of $160,000) and to begin to archive television news broadcasts began operating at the end of the year.

9.7 2000: Building Thematic Web Collections

By 2000, the Archive had achieved a level of technical stability. Acceptance of crawling data was routine, and the migration from tape to disk was long over.

Table 9.1. The Archive's Internet Collections as of March, 2000

Collection	Units	Size
Web (1996 to 3/2000)	1 billion pages	13.8 terabytes (TB)
FTP (1996)	50,000 sites	.05 TB
Usenet (1996-1998)	16 million postings	.592 TB

2000 was another election year in the US, and this time most of the electorate had Internet Access. It was clear to the political establishment that a presence on the Internet was vital to winning the election, and with this increased focus on politics online, the Internet Archive partnered with the Library of Congress to collect political sites.

This was the Archive's first project with the Library of Congress, and for many on staff at the Archive, marked a transition from experimental project into an established institution.

[2] http://netpreserve.org

The idea of providing access to preserved ephemeral works gained momentum, and the Moving Images Archive was released in 2000. Now under Creative Commons licenses, the Moving Images Archive consists primarily of films from the Prelinger Archive, a collection of over 1,900 "ephemeral" (advertising, educational, industrial, and amateur) films. The collection currently contains over 10% of the total production of ephemeral films between 1927 and 1987 in the USA, and is one of the most complete and varied collection in existence of films from these poorly preserved genres.

9.8 2001: Public Access with the Wayback Machine: The 9/11 Archive

Between March, 2000 and March, 2001, the Archive tripled the size of its holdings to a total of more than 40 TB. At that point, the Archive was growing by roughly 10 TB per month.

Table 9.2. The Archive's Collections as of March 2001

Collection	Units	Size
1996 – 3/2001	4 billion pages	40 TB
Election 2000 Archive	200 million pages	2 TB
Usenet: 1996 – 1998, 2000 – 3/2001	16 million postings	.5 TB
Archival movies: ca. 1903 to ca. 1973	360 movies	.5 TB
Arpanet: Historical documentation	5,000 pages	< .1 TB

But 2001 was a difficult year for many high tech organizations in the San Francisco area. The collapse of the stock market, the demise of hundreds of local companies, and the World Trade Center attack in New York all had an effect on the Archive's operations. In particular, the loss of high tech jobs in the San Francisco area made it easier to hire engineers, and archiving the events 9/11 provided a focus, as well as a test of the Archive's ability to handle moving images and to respond to events.

In early 2001, perhaps the main question facing the Archive was how best to provide access to the collection. Some data was served directly to the general public via the Alexa service, but direct access to the collections still required Unix programming skills.

Working under contract to the Archive, programmers at Alexa built the Wayback Machine, which serves up contents of the Archive based on URLs. On October 24, 2001, the Wayback Machine went live, offering access to more than 10 billion archived Web pages and 100 TB of data.

At that time, data was stored on Hewlett Packard and uslab.com servers running the FreeBSD and Linux operating systems. Each computer had about 512 Mb of memory and generally held just over 300 GB of data on IDE disks.

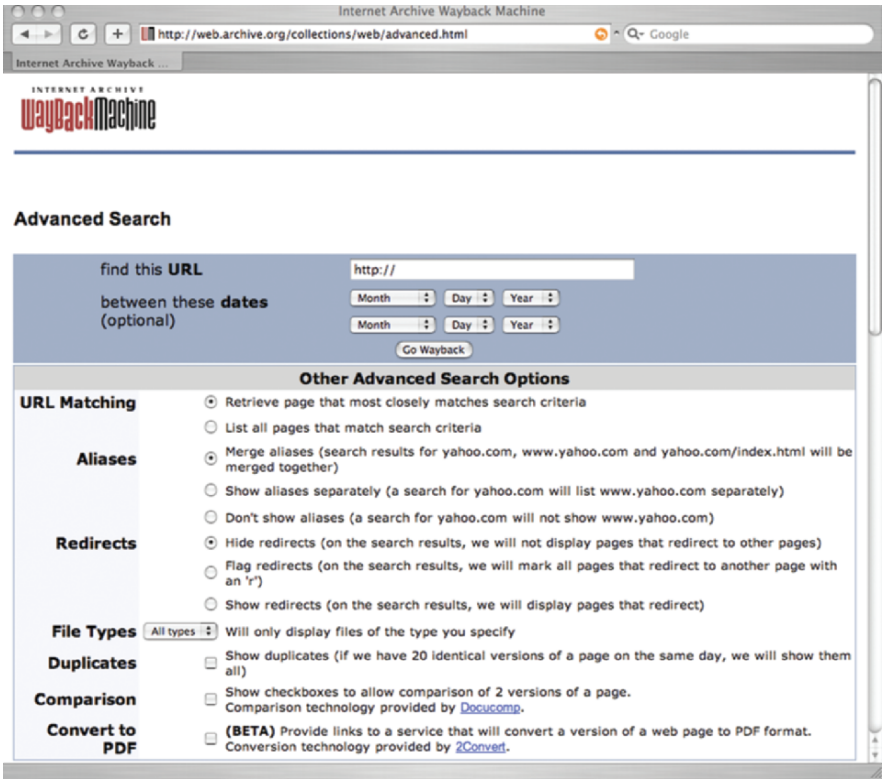


Fig. 9.1. Current search options for the Wayback machine

The other major project of 2001 was the September 11 Archive. Working with the Library of Congress, the Archive collected Images from over 30,000 selected websites from September 11, 2001 to December 1, 2001, and hundreds of hours of broadcast news footage.

9.9 2002: The Library of Alexandria, The Bookmobile, and Copyrights

The Archive undertook five major projects in 2002 to expand the breadth of its collections, the options for accessing those collections, its partnerships with other organizations, and its role in policymaking.

The first and largest was the creation of a mirror site at the Library of Alexandria in Alexandria, Egypt. Servers and more than 100 TB of data, valued at more than $5 million, were shipped to Egypt and installed in time for the Library's grand opening in April.

The second major project was the creation of the Internet Bookmobile, designed to show how the combination of electronic scans of books, print on demand technology and a satellite network connection could effectively fit a million-book library in the back of a van. Partnering with Carnegie Mellon in the summer of 2002, the Million Books Project (MBP) was launched, aiming to digitize at least one million books and offer them free-to-read on the Internet. With encouragement from Suzanne Mubarak, the Archive began building a bookmobile in the U.S., and working with others in India and Kenya to clone its basic prototype.

The third major project was policy related. On September 30, 2002, in an effort to raise public awareness of important copyright policy issues, Internet Archive's Bookmobile embarked on a cross-country journey to print and deliver free books. The Bookmobile parked and printed books at the United States Supreme Court building where, on October 9, the Justices heard arguments in Eldred vs. Ashcroft, a landmark case that decided how many books would be part of theBookmobile's digital library and all other digital libraries in the US Unfortunately, Eldred was defeated and the copyright extension went into effect, but the Bookmobile project blossomed and eventually became its own non-profit. Currently the Government of India is building 25 bookmobiles for use throughout India.

The fourth area of activity involved the creation of the Archive's first book and music collections. In June, the first book collections were put online; in August, the Live Music Archive, a collection of concert performances that could be downloaded legally, went online.

The fifth major project was the launch of the International Children's Digital Library, in partnership with the University of Maryland, and supported by Library of Congress, NSF, IMLS, Kahle/Austin Foundation, Adobe Systems Inc., the Markle Foundation, and Octavo. The ICDL was and is focused on the inherent promise of the Internet to provide direct and global access to quality content for children.

At the end of 2002, the Archive led an effort to ensure the integrity of digital archives by standardizing the criteria under which materials might be removed or made unaccessible. In a meeting at UC Berkeley, representatives from the Archive met with other digital librarians to evolve the Oakland Archive Policy (Ubois 2002),[3] which details procedures for disclosing removals of materials as required by law, or as requested by site owners and others (see also Chapter 1 (Masanès 2006)).

[3] See see http://www.sims.berkeley.edu/research/conferences/aps/removal-policy.html

9.10 2003: Extending Our Reach via National Libraries and Educational Institutions

In 2003, the Archive continued to reach out to national libraries and educational institutions around the world. With the International Internet Preservation Consortium (IIPC), the Archive began working closely with participating organizations on new standards and a new open source crawler.

In July, 2003, the Archive helped to launch the International Internet Preservation Consortium, a group of 12 national libraries that agreed to work on developing standards, tools and policies to to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations. To achieve this mission, the IIPC is working to accomplish the following goals:

- To enable the collection of a rich body of Internet content from around the world to be preserved in a way that it can be archived, secured and accessed over time;
- To foster the development and use of common tools, techniques and standards that enable the creation of international archives;
- To encourage and support national libraries everywhere to address Internet archiving and preservation.

The IIPC was chartered at the Bibliothèque nationale de France with 12 participating institutions. The members agreed jointly to fund and participate in projects and working groups to accomplish the goals of the IIPC. The initial agreement is in effect for 3 years. During that period, the membership is limited to charter institutions.

In 2003, the Archive received significant outside funding from outside organizations, including, the Hewlett and Sloan Foundations, and began work on a series of special collections.

Further declines in the cost of disk storage and internet bandwidth led the Archive to make a standing offer of "unlimited bandwidth, forever, for free" to organizations and individuals with digital materials.

This offer led to a partnership with Etree, an all-volunteer organization founded in 1998 to enable free, legal trading of recordings of live music concerts. As a result of the partnership with Etree, the Archive now hosts more than 15,000 live music concerts.

To support growing needs for both storage and bandwidth, the Archive opened a new data center in San Francisco. The new data center was connected to the Internet with a 1 Gbps link and housed more than 1,500 commodity PCs all running Linux.

9.11 2004: And the European Archive and the Petabox

In 2004, the Archive began migrating data to its third generation of hardware, known as the Petabox. Based on rack-mounted commodity hardware and the Linux operating system, the Petabox design offered RAID storage for roughly $2,000 per terabyte, or $2 million per petabyte.

The first installation of the new design was in Amsterdam, at the newly formed European Archive, an institution intended to serve the needs of the European community that Internet Archive supported with other european partners in its first year of existence. The installation in Amsterdam is growing to provide a mirror to the collections in Alexandria and San Francisco. Creating a network of independent institutions around the world, each capable of operating independently, will help to prevent catastrophic losses of information.

Also in 2004, the International Internet Preservation Consortium launched Heretrix, an open-source, extensible, Web-scale, archival-quality Java-based Web crawler. Now in use by many heritage institutions worldwide, Heretrix supports multiple different use cases including focused and broad crawling.

Collection development in 2004 took major steps forward through the hiring of additional staff, completion of book and movie scanning projects, donations of data from other institutions.

9.12 The Future

Advances in computing and communications make it possible to cost-effectively store every book, sound recording, movie, software package, and public Web page ever created, and provide access to these collections via the Internet to students and adults all over the world.

As stated in the Universal Declaration of Human Rights, Article 19 “Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media regardless of frontiers.”

For the years ahead, the Archive’s mission is clear: to create a new Universal Library that makes all knowledge easily available to every man woman and child around the world.

References

Burner, M. & Kahle, B. (1996). The ARC File Format

Gray, J. & Shenoy, P. (1999). Rules of Thumb in Data Engineering. *Microsoft Technical Report*, *MS-TR-99-100*

Masanès, J. (2006). Web archiving: issues and methods. In J. Masanès (Ed.), *Web Archiving*. Springer, Berlin Heidelberg New York

Ubois, J. (2002). The Oakland Archive Policy. Recommendations for Managing Removal Requests And Preserving Archival Integrity

10 Small Scale Academic Web Archiving: DACHS

Hanno E. Lecher

Leiden University
h.e.lecher@let.leidenuniv.nl

10.1 Why Small Scale Academic Archiving?

Considering the complexities of Web archiving and the demands on hard- and software as well as on expertise and personnel, one wonders whether such projects are only feasible for large scale institutions such as national libraries, or whether smaller institutions such as museums, university departments and the like would also be able to perform the tasks required for a Web archive with long-term perspective.

Even if the answer to this is yes, the question remains whether this is necessary at all. One could think that the Internet Archive in combination with the efforts of the increasing number of national libraries is already covering many, if not most relevant Web resources. Does academic or other small scale Web archiving make sense at all?

Let me begin with this second question. The Internet Archive has done groundbreaking work as the first initiative attempting comprehensive archiving of Web resources. Its success in accomplishing this has been revolutionary, and it has laid the foundation on which many other projects have built their work. Still, examining what the Internet Archive and other holistic projects[1] can achieve it is easy to discover some limitations. Since their focus of collection is very broad, they have to rely on robots for a large part of their collecting activities, automatically grabbing as many Web pages as possible. This kind of capturing is often very superficial, missing parts located further down the tree, many pages being downloaded incompletely, and some file types as well as the hidden Web being ignored altogether.

In addition, since harvesting is performed automatically and in irregular intervals there can be no conscious selection of resources, and no possibility

[1] In contrast to topical archives, I consider holistic Web archives to try to capture either the contents of the whole Internet, or at least certain domains thereof, using automatic crawling routines.

to consider or detect important contents whose existence might be very short lived or difficult to detect. Of course, the Internet Archive and other large scale Web archiving projects do feature special collections, where much effort is spent on developing "deep" collections around a chosen number of topics. The number of these topics, however, is very limited, and it is obvious that many research projects will have to develop their own archives.

A further difficulty is the restricted accessibility to the contents of holistic Web archiving projects. Currently the exact (former) URL of a document or website has to be known to be able to retrieve the data. Usually it is not possible to search these archives using descriptive metadata or full-text indexing services. Even if a full text search option was available, the lack of conscious resource selection would only provide a similar amorphous result as a Web search today.

Looking at these limitations it becomes evident that the Internet Archive and others are neither archiving the Internet in its completeness nor provide ways of access suitable for many academic or other research purposes. Small scale academic archiving thus becomes an important need. But, returning to the question raised in the beginning is this feasible?

There are some issues to be approached differently or to be considered if the organization planning a Web archive is of smaller scope and has only very limited human and/or hardware resources. To illustrate some of the main issues, the Digital Archive for Chinese Studies (DACHS) will be examined as a case study. It should be kept in mind, though, that different projects will frequently need different approaches or solutions.

10.2 Digital Archive for Chinese Studies

The main objectives of the DACHS[2] are to identify and archive Internet resources relevant for Chinese Studies in order to ensure their long-term accessibility. Selection plays an important role in this process, and special emphasis is put on social and political discourse as reflected by articulations on the Chinese Internet.

Currently the DACHS project is handled by the libraries of two Sinological institutes, namely, the Institute of Chinese Studies at Heidelberg University, Germany, and the Sinological Institute at Leiden University in the Netherlands. The organizational infrastructure is thus quite different from big national libraries (see Table 10.1).

[2] Accessible at http://www.sino.uni-heidelberg.de/dachs/

Table 10.1. DACHS collection holdings (2005)

What	Number of files	Size in GB
Discussion boards	2,04,349	1.6
Documents	2,35,449	4.3
Donations	979	0.15
Films	982	0.427
Journals & Newsletters	2,62,939	5.8
Websites	12,65,857	24
Total	19,70,555	36.277

When the idea of downloading online resources from or about China was first brought up in late 1999 in Heidelberg, it was by no means clear what this would mean. The still much undeveloped idea was introduced as a possible part of larger application for creating a European Center for Digital Resources in Chinese Studies aiming at the improvement of the conditions for China-related research and information access in Europe. The project included a number of activities such as purchasing a wide range of commercial full-text databases, supporting the development of academic database projects as well as developing own ones, developing finding aids for printed and non-printed resources on China, and, above all, providing free access to all resources as widely as possible.[3] The project was granted a term of five years, including the financial means to improve an already existing hardware environment and to hire some personnel in the form of student assistants.

The main guideline of the project was put as "maximum flexibility with maximum accountability". This simply meant that there was room to develop detailed plans for many sub-projects on the go and even ideas for new sub-projects as needed.

While the year 2000 was spent to get most of the sub-projects in more advanced planning status running and to reorganize the IT infrastructure of the institute, concrete planning for the Web archive only began in 2001.

10.2.1 Initial Steps

Looking at the existing infrastructure from which the Web archive had to be developed, it became very obvious that the limits were tight indeed. The Digital Archive for Chinese Studies was to be run by the library of the

[3] For a detailed description of the European Center for Digital Resources in Chinese Studies and its sub-projects please visit the project's homepage at http://www.chinaresource.org/

institute[4] which at the same time also oversaw the IT environment with its then four servers[5] and close to 100 workstations. Responsible for maintaining this IT infrastructure was the librarian with the help of one or two part-time student assistants and – if needed – support from the ICT department of the University.

For the Web archiving project an additional part-time student assistant could be hired for handling the actual workflow (downloading, archiving, metadata creation, etc.), while the librarian – in addition to his responsibilities for the library and the IT environment – would take care of project management and conceptual development. In fact, the contribution of the project assistant to the conceptual development of DACHS was substantial. Involving the assistant to a very high degree was considered important to avoid concentration of knowledge about the theoretical framework of the project in the hands of only one person. This strategy paid off when the project management was seamlessly passed on after the librarian left the institute for another work place.[6]

Of course a number of issues had to be considered at the beginning of the project, taking into account the size of the institute and its possibilities. What would it mean to aim at long-term accessibility of archived resources? What were the requirements on hardware and software to create and maintain such an archive? How the selection of resources should be organized as an ongoing task, and how should the data be made accessible? And above of all: what else needed to be considered for proper planning, and where to look for answers?

Answers to the last question could be found in a document that in 2003 was to become the standard framework for Internet archiving, canonized as ISO 14721:2003, but generally known as OAIS, the Open Archival Information System.[7] This document proved to be of crucial importance to our project since it provided us with much needed theoretical background and helped us to pinpoint most of the critical issues of Web archiving.

---

[4] In 2001 the library of the Institute of Chinese Studies was home to some 1,10,000 volumes and staffed with one full-time librarian and about 20 part-time student assistants.

[5] Two Sun Sparc stations running various online full text as well as bibliographic databases, a Linux machine serving our WWW pages and e-mail services, and a Novell server providing the fileserver infrastructure for the institute.

[6] The librarian accepted the position as head librarian of the Institute of Chinese Studies at Leiden University. It is thus no accident that Leiden later became the first new project partner of DACHS.

[7] The document is available at http://www.ccsds.org/documents/650x0b1.pdf . Cf. also Chap. 8 of this book.

Useful as this document is, it has one major drawback: as its name suggests it is a framework only, giving but very theoretical guidance on the various issues and leaving the concrete implementation to the ingenuity of the user. It became thus necessary to look for other places in order to come to an understanding of how to put the framework into practice. Much information could be found at places such as PADI, RLG and others,[8] but even more important was the active participation in workshops and conferences that were dealing with issues related to Web archiving.

10.2.2 Institutional Sustainability

One of the most important questions the whole project had to face was how to provide institutional sustainability for such a long-term archiving project. None of the three major factors that were decisive in this question could be taken for granted on a long-term basis: financing of the project could run out; interests of people in the institute could shift, leading to negligence of the project; and even the long-term status of the institute itself was by no means guaranteed. It was thus obvious that an institute represented a far less trustable place for long-term archiving than a national library or archive, where legal provisions forced the institution to fulfil its responsibility towards the collection basically forever.

Strategies had thus to be developed to make the survival of the archive possible even after the institute itself ceased to exist or was not able anymore to support the project. Survival could be defined in two ways: either it should be possible to keep the archive active, which means that all activities from resource selection to ingest to making the data accessible would continue; or the archive should at least be preserved in a deep frozen status, which means that although no new data come into the archive, at least the accessibility of the already existing ones should be ensured.

The way to accomplish this is again twofold. First and foremost, the archive has to fulfil the basic attributes of a trusted repository, which means adherence to acknowledged standards as described in the OAIS model. This would make it possible for other institutions – most ideally a national library – to take over the contents should we not be able to accommodate the archive anymore.

The second way was to develop the archive into a distributed effort. If a number of institutions were actively and interactively participating in the project it would be possible for one partner to take over an archive from another institution that had to discontinue. Here again the adherence to

[8] Cf. list of resources at the end of this chapter.

established standards were essential. For our project it was decided to work towards a collaborative project and to look for possible partners as soon as most of the local issues were solved and the archive had won recognition as an essential contribution to the field of Chinese Studies.

10.2.3 Hardware

To ensure the proper working of DACHS on the local level a suitable hardware environment had to be set up. Next to providing scaleable server space and dedicated workstations for the download routines and management purposes we had to give special consideration to backup facilities, fall-out security and virus protection.

The computer center of Heidelberg University provides a sophisticated backup system using the IBM ADSTAR Distributed Storage Manager (ADSM). Using this system a backup of the whole Archive is made every night onto magnetic tapes stored at the computer center. Regular backup copies of these tapes are also stored at the University of Karlsruhe. The data of the archive are thus kept at three different places which offer reasonable security.

An Uninterruptible Power Supply (UPS) device as well as a RAID (Redundant Array of Inexpensive/ Independent Disks) system (level one) was installed to provide basic security for uninterrupted availability. For virus protection, we turned to McAfee Virus Scan software. Using virus definitions fetched on an hourly basis from the McAfee server, all incoming data are routinely checked on viruses, and cron jobs automatically incite regular scan processes of the whole archive.

Basically we could make use of an already existing IT infrastructure at our institute which also included backing by the University ICT department for the more intricate parts. Much of the above should be standard equipment for universities anyway, mostly however, not within an institute but rather maintained at central ICT facilities. This can be an advantage since a more professional environment can be counted on. It may also be a handicap, though, if restrictions on hard- and software apply.

10.2.4 Software

In order to actually start with the Digital Archive for Chinese Studies project a piece of software was now needed that had to fulfil a number of conditions.

The Internet can be seen as a huge collection of interlinked data in various formats and encodings. Archiving such data means to preserve them in

a way that content, functionality and look and feel are kept as close to the original as possible. Since look and feel of Web content does vary depending on the browsing software there is no way to preserve this in a reliable way. However, it was felt that if we could preserve the original bit stream most of the requirements above should be met. When downloading whole websites, the original file structure should be kept intact. Of course, to keep linking between the downloaded documents intact all links had to be converted into relative links on the fly. We needed thus configurable crawling software that was able to perform this task. The software should be able to handle a wide range of different formats, including dynamically generated contents of billboards and the like. And it should be affordable. After some testing we found that Offline Explorer Pro by MetaProducts provided all needed functionality, at least for the beginning.

10.2.5 Metadata

An issue that did engage us considerably was the creation of metadata. On the one hand metadata offer an important access point for users since they provide harmonized information describing the content of the document, such as author, title, and subject. This kind of access is particularly important as long as no full text search can be provided. At the same time, technical metadata are unanimously described as essential to assist in long-term preservation since they have to carry all sorts of information necessary for proper administration and future handling.

Metadata are very costly to create, though, since even with semi-automated metadata harvesting routines much of the work has to be done by hand. Very obviously, it is impossible to create metadata for hundreds of thousands of documents at the rate they are downloaded from the Internet (see Fig. 10.1). And even if this were possible it proved to be very hard to find out what exactly these metadata should contain, to what degree of granularity, and in which format.

One of the questions that arose from these considerations was whether or not metadata were necessary at all. Or better: would it be possible to solely rely on search algorithms to retrieve all data needed for access as well as for long-term preservation purposes? From the user's perspective it was argued that full text search was a much more reliable tool for finding documents than crude subject headings or title words contained in often imperfect metadata. From the technical point of view much of the necessary information such as file format, download date, encoding, etc. should easily be retrievable from the data themselves, or else could be made part of the file naming structure.

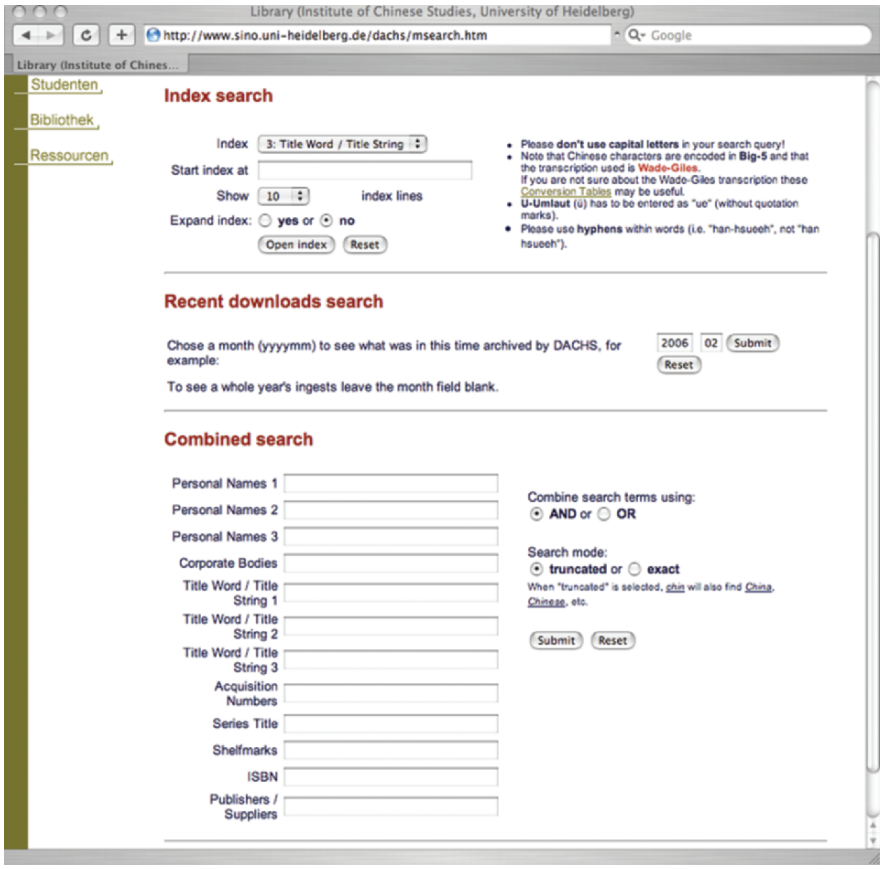


Fig. 10.1. DACHS metadata search interface that enables advance search on the archive

In the end this idea was rejected. Digital resources are an extremely volatile form of information, and separate documentation on these resources is necessary to ensure their integrity and accessibility for the long-term future. Depending on the data themselves for this kind of information makes them extremely vulnerable and prone to rapid decay.[9]

To make the task manageable and to better integrate the archived material into the holdings of the library at large it was decided to create metadata as part of our regular catalogue. The catalogue was thus adapted to

[9] Discussion and a framework of preservation metadata for digital resources can be found at the homepage of the joint OCLC/RLG project "PREMIS – PREservation Metadata: Implementation Strategies",http://www.oclc.org/research/projects/pmwg/.

accommodate the necessary additional metadata, including categories for rights management, history of origin, management history, file types, identifiers, and others. Depending on the complexity of the resource we now create metadata records that may either describe a single file, as in the case of single text-only documents, or a whole set of files if the archived object is a website, discussion board or newspaper.

10.2.6 Collection Policy

Quite obviously the aim of the Digital Archive for Chinese Studies cannot be to preserve the Chinese Internet in its entirety. Neither is this technically an option, nor do we think this to be very useful. As a research institute, we are interested in parts of the Chinese Internet that reflect certain aspects of the Chinese society, and that are particularly ephemeral. Providing an informed selection of resources is thus an asset that helps current as well as future users to detect material we believe to be relevant. Of course values will change over time, and later generations might well have preferred different choices. In this case the Internet Archive still provides a very rich alternative where other choices will be available. It is the combination of the two approaches – selective and holistic – that provides the future user with the widest array of possibilities for his or her research.

At the same time, however, we do want to reduce the danger of being too narrow in our selection. As will be discussed below, DACHS is evolving into a larger cooperative project. Different partners will have different selection criteria and thus apply different sets of values to their selection process. This can only enrich the contents of the archive while the policy of informed resource selection remains intact.

Still, in order to make best use of the limited resources available to us we have to develop clever collection strategies. Given the number of Internet sites in or about China, and the pace in which relevant discussions begin, develop and disappear on the Chinese Net, the task of building a meaningful archive of selected online resources is a daunting one. To cope with this situation we started to build an information network of individuals (native or foreign scholars and *"netizens"*) who are actively or passively part of the discourse we try to grasp. Making use of their judgment and knowledge it becomes possible to identify specific sites or discussion processes that fit into our selection profile. The larger this information network the more diversified this part of the selection process can be.

Usually the information network also serves another purpose. Since its members are part of the contemporary Internet culture they can provide valuable context information about the archived resources. As much as

possible of this context information is created and preserved either in the metadata themselves or on dedicated Web pages that form part of the collection.

Since a few months we are also working on a number of special projects such as contemporary poetry, SARS, or the homosexual scene in China. Scholars and MA students working on these topics try to create comprehensive archives, including endangered, significant or representative websites, as well as other material such as photos and posters. To put these resources into context they add introductory texts as well as their own research material to the archive.[10] We consider this kind of value added approach with its contemporary contextualization as essential for a better understanding of these volatile resources especially in the distant future.

But we do not solely rely on informants and scholars. Certain events of international impact such as the September 11 terror attack or the Olympic Games in China in 2008 frequently cause heated discussions on various platforms on the Internet. To capture such outbreaks of public opinion we are working through checklists of relevant discussion boards and newspapers, resulting in a set of snapshots of relevant material covering a time span of a few weeks before and after such events.

Somewhat similar to the special projects described above are collections donated or sold to DACHS by private persons, researchers, research groups, or other institutions. Sometimes, we are even approached by the publishers of endangered websites (or we contact them) to help them preserve the content. These collections where not specifically created for DACHS (and may thus not always follow all quality standards), but DACHS helps to keep these resources available on a long-term basis.

10.2.7 Partnerships

We have seen earlier that running the archive as a distributed collaborative effort involving coequal partners is an essential surviving strategy for small scale Web archives. Partnerships may help to ensure long-term preservation goals through offering inheritance service for members that have to terminate operations. They also allow for a distribution of workload, thus providing means of cost reduction and the possibility of a wider selection of archived resources. Hardware, experience and quality standards can also be meaningfully shared and might improve the overall performance of the archive. Not the least important are partnerships as part of a political

[10] See e.g., the poetry section of the DACHS Leiden division at http://www.sino.uni-heidelberg.de/dachs/leiden/poetry/ as an example.

strategy. Being an internationally collaborative project it is definitely easier to gain recognition from the scholarly community and to successfully campaign for funding.

Recognizing these issues DACHS has begun to develop guidelines for possible partnerships. These guidelines are being explored together with Leiden University in the Netherlands, which is participating in the project since the end of 2003. The baseline assumption is that partners should be able to remain as independent as possible and keep their own identity, while there are a number of standards and services that need to be shared.

One of our major concerns is to build common finding aids, including a hyperlinked subject guide (or table of content) as well as a full text and a metadata search facility integrating the complete archive with all its current and future project partners. This not only requires a central access point in the form of a joint home page, but also considerations about how these search options need to be constructed to serve the desired purpose. Full text search must be possible across different domains, and for the metadata either a union catalogue is needed where all data are physically stored together, or a virtual catalogue has to be constructed that searches the various local metadata and presents them in a coherent manner. In any case there is a strong need to establish shared standards for the creation of the metadata.

Other issues that need to be discussed include a common access restriction policy, regulations for the division of labour, and routines that help to avoid duplication of efforts when archiving resources more than once.

Important for the design of the cooperation is also the degree to which prospective partners are able to either build a fully-fledged local archive infrastructure, or to make use of the facilities of one of the project partners. Since all the above issues are still work in progress it is impossible to share experiences at this point of time. It is well possible that more issues will emerge in the process of negotiation and actually implementing the partnership, but it has become quite clear that for small scale Web archiving projects cooperation on many levels is essential.

10.3 Lessons Learned: Summing Up

After having built DACHS for about four years now, many questions and problems still remain unresolved or in development. Looking back, however, there are a few issues we would try to deal with differently today than we did in the beginning.

Most essential in this respect is the allocation of positions for the project. In addition to the daily working routines of data ingest, the task of developing and managing a Web archiving project must not be underestimated. The decision to put this task exclusively into the hands of a librarian who is already performing a plethora of other duties is at least questionable. While his involvement into the development of the project has many advantages and important, a dedicated managerial position for DACHS would have been is much better suited to lead the project into smooth operation. Many issues could have been addressed much earlier and more effectively, thus positively effecting the project's development.

A second issue that – at least from today's perspective – should be solved differently is the choice of harvesting and archiving software. At the time DACHS started operation in 2001 Web archiving was still in its infancy, and many big players of today only just started to develop and publish their efforts. Although the software we chose for our purpose at the time fulfilled most of our needs to our satisfaction, the tools available by now are much better suited for the task. Without going into the details of possible candidates – this is done elsewhere in this book – it remains clear that a new choice will have to be made for our project in order to streamline the ingest process and creation of metadata.

The above may also illustrate a final point I would like to make. Many issues that make Web archiving difficult and in times even discouraging need not be solved by yourself. There is no need to worry about building a file format repository or developing your own harvesting software. It is possible – even necessary – to rely on the efforts of others. Collaboration does not only include partners of your archiving project, but also colleagues and institutions of the Web archiving community providing solutions and tools that you would not be able to create if on your own.

10.4 Useful Resources

The following list of resources is only a very small selection of what we found useful for our work. Many more resources will be found if you visit the websites below, especially the excellent subject gateway to digital preservation, PADI. All resources were accessible in May 2006.

10.4.1 Websites

Council on Library and Information Resources (CLIR)
http://www.clir.org/
Electronic Resource Preservation and Access Network (erpaNet)
http://www.erpanet.org/
International Internet Preservation Consortium
http://netpreserve.org/
Internet Archive
http://www.archive.org/
Networked European Deposit Library (NedLib)
http://www.kb.nl/coop/nedlib/
PADI Preserving Access to Digital Information (National Library of Australia)
http://www.nla.gov.au/padi/
PADI: Web archiving
http://www.nla.gov.au/padi/topics/92.html

10.4.2 Mailing Lists

Archivists
http://groups.yahoo.com/group/archivists/
DigiCULT
http://www.digicult.info/pages/subscribe.php
DIGLIB – Digital Libraries Research mailing list (IFLA)
http://infoserv.inist.fr/wwsympa.fcgi/info/diglib/
OAIS Implementers (RLG)
http://lists2.rlg.org/cgi-bin/lyris.pl?enter=oais-implementers
PadiForum (National Library of Australia)
http://www.nla.gov.au/padi/forum/
Web-Archive
http://listes.cru.fr/wws/info/web-archive/

10.4.3 Newsletters and Magazines

CLIR Issues
http://www.clir.org/pubs/issues/
DigiCULT Newsletter
http://www.digicult.info/pages/newsletter.php
D-Lib Magazine
http://www.dlib.org/
DPC/PADI What is new in digital preservation
http://www.nla.gov.au/padi/qdigest/
RLG DigiNews
http://www.rlg.org/en/page.php?Page_ID=12081

List of Acronyms

AIP	Archival Information Package
AOLAP	Austrian On-Line Archive Processing
API	Application Programming Interface
CAIDA	Cooperative Association for Internet Data Analysis
CCSDS	Consultative Committee for Space Data Systems
CGI	Common Gateway Interface
CMS	Content Management System
CPU	Central Processing Unit
DACHS	Digital Archive for Chinese Studies
DAVID	Digitale Archivering in Vlaamse Instellingen en Diensten
DB	Database
DCC	Digital Curation Center
DIP	Dissemination Information Package
DMCA	Digital Millennium Copyright Act
DNS	Domain Name System
DOI	Digital Object Identifier
DSL	Digital Subscriber Line
DTD	Document Type Definition
DWH	Data Wharehouse
ECMA	European Computer Manufacturer's Association
FTP	File Transfer Protocol
GIF	Graphics Interchange Format
GUI	Graphical User Interface
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ICANN	Internet Corporation for Assigned Names and Numbers
IIPC	International Internet Preservation Consortium
IRC	Internet Relay Chat
IWAW	International Web Archiving Workshop
JHOVE	JSTOR/Harvard Object Validation Environment
JPEG	Joint Photographic Experts Group
MARC	Machine-Readable Cataloging
MIME	Multipurpose Internet Mail Extensions
MINERVA	Mapping the INternet Electronic Resources
NEDLIB	Networked European Deposit Library
NUTCH-WAX	Nutch + Web Archive eXtensions

NWA	Nordic Web Archive
OAI	Open Archives Initiative
OAIS	Open Archival Information System
OCLC	Online Computer Library Center
PADI	Preserving Access to Digital Information
PANDORA	Preserving and Accessing Networked Documentary Resources of Australia
PDI	Preservation Description Information
PIW	Publicly Indexable Web
PREMIS	PREservation Metadata Implementation Strategies
RAID	Redundant Array of Independent Disks
RDF	Resource Description Framework
RFC	Request For Comments
RIPE	Réseaux IP Européens
RLG	Research Libraries Group
RSS	Really Simple Syndication
SIP	Submission Information Package
SLD	Second-Level Domain
TCP	Transmission Control Protocol
TIFF	Tagged Image File Format
TLD	Top-Level Domain
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UVC	Universal Virtual Computer
W3C	World Wide Web Consortium
WA	Web Archive
WAIS	Wide Area Information Servers
WARC	Web ARChive file format
WERA	WEb aRchive Access
WIS	Web Information Systems
XML	eXtensible Markup Language

Index