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 back-end 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

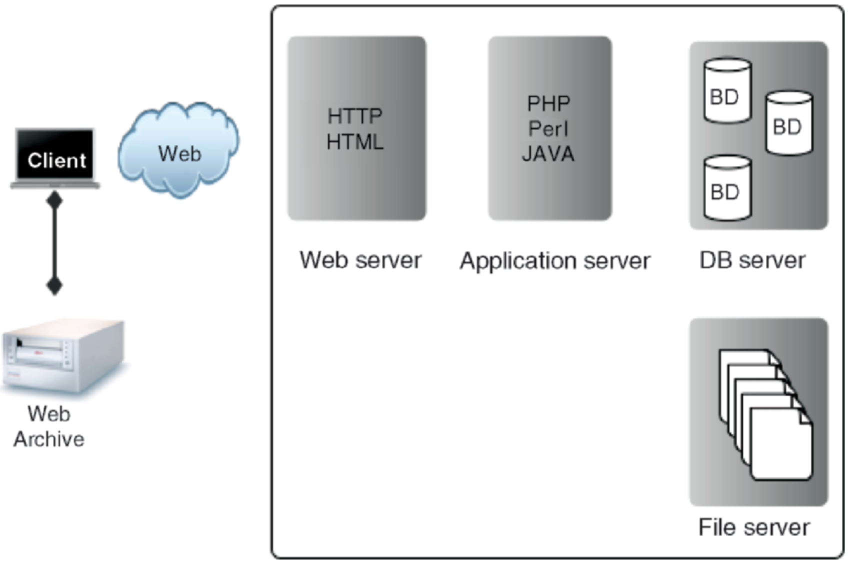


Fig. 1.1. Client-side archiving: the Web archive is in position of client to gather content from the Web server. The Web server can generate content from various other servers (application, database, file servers)

[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

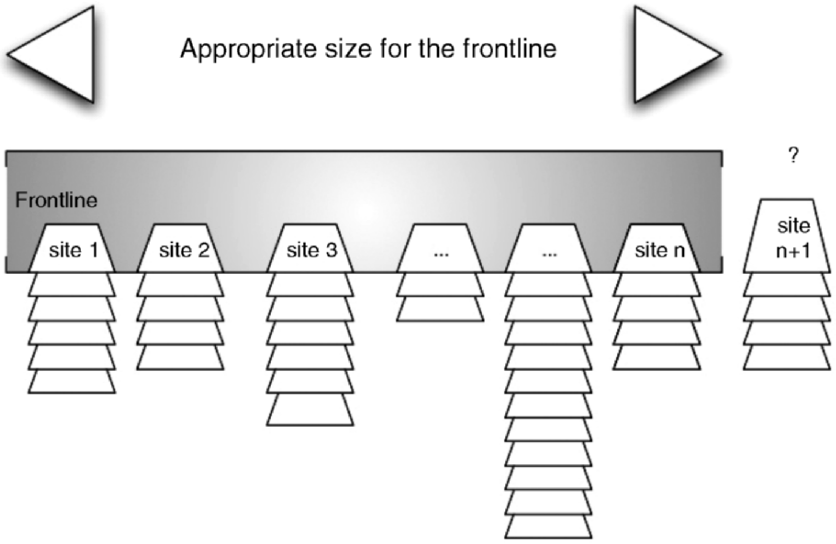
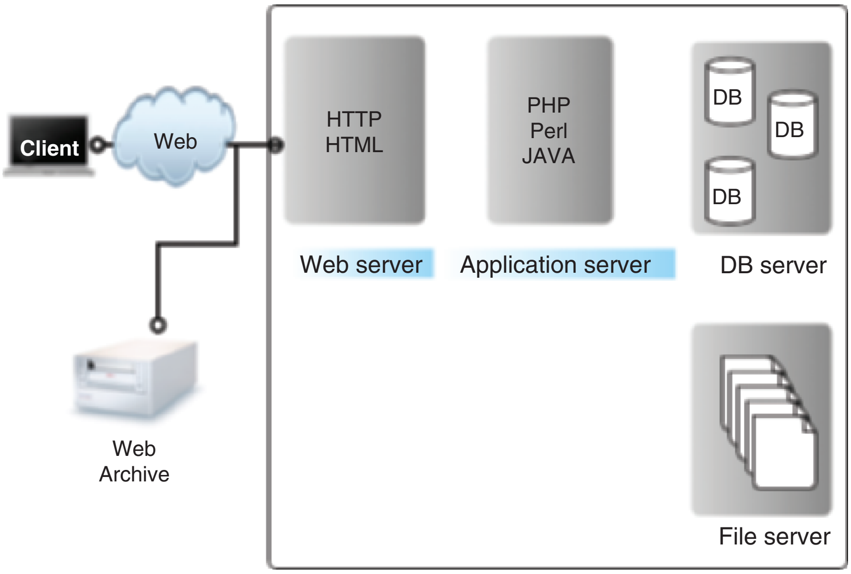


Fig. 1.2. The frontline contains sites to be crawled at the same time by the same crawling infrastructure. Size of the frontline (*n*) is optimized if delay between requests is limited by politeness rules only and crawling resources are kept busy. If *n* + 1 sites were crawled, crawling resources limitation 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 Webserver'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

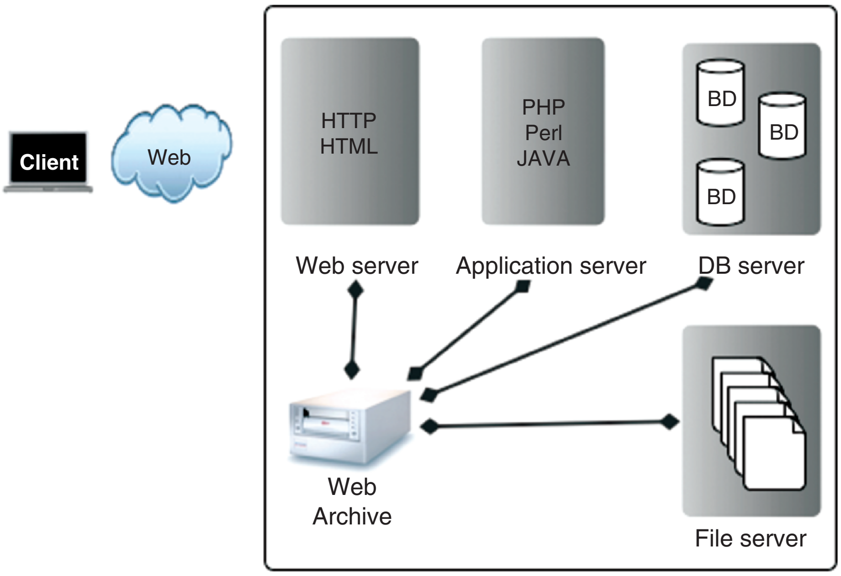


Fig. 1.4. Server-side archiving: different pieces of information are archived directly from servers. Generating a working version of the archived content and not only a back-up of files is the challenge of this method

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.

<a href=“example.org/example.HTML ”> </a>

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.

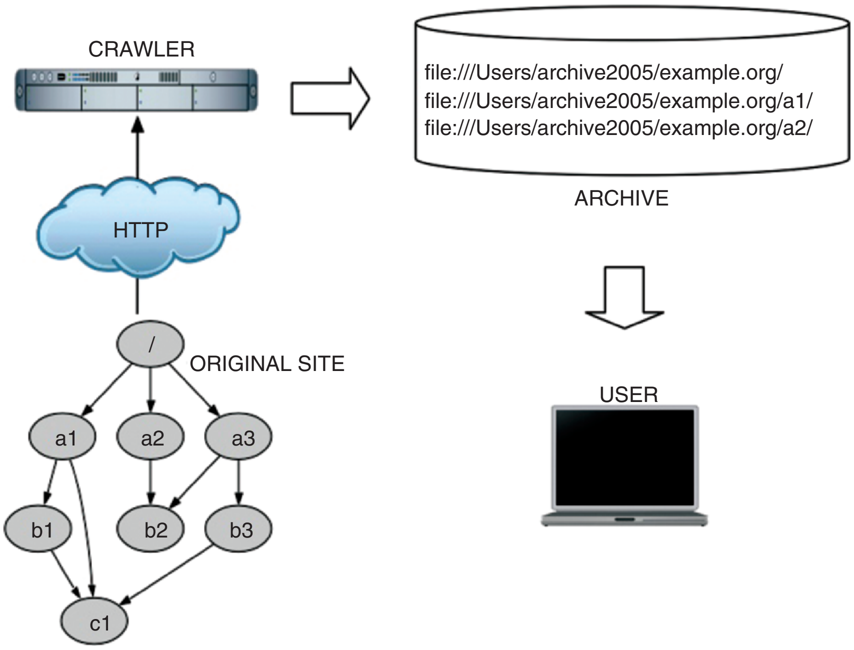

CRAWLER
file:///Users/archive2005/example.org/
file:///Users/archive2005/example.org/a1/
file:///Users/archive2005/example.org/a2/
ARCHIVE
HTTP
/
ORIGINAL SITE
a1
a2
a3
b1
b2
b3
c1
USER

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.

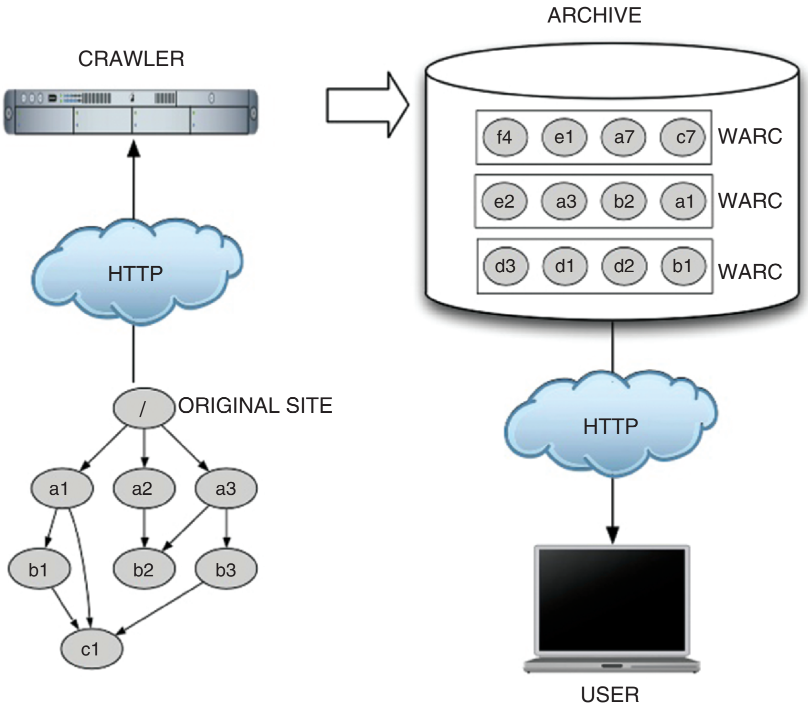


Fig. 1.6. The Web-served model: the original site is crawled and responses are stored unchanged in container (WARC files) which permits to avoid mapping to the file system's naming conventions and changing link structure. Access requires a Web server that fetches content in the containers and sends it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

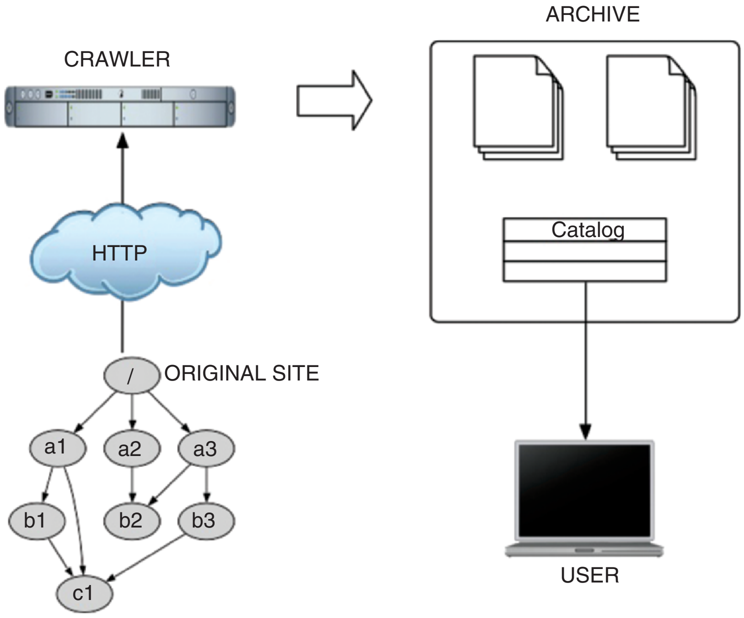


Fig. 1.7. Documents from the original site are re-organized in the archive, following a non-Web structure, using for instance a catalogue providing access to individual documents

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).

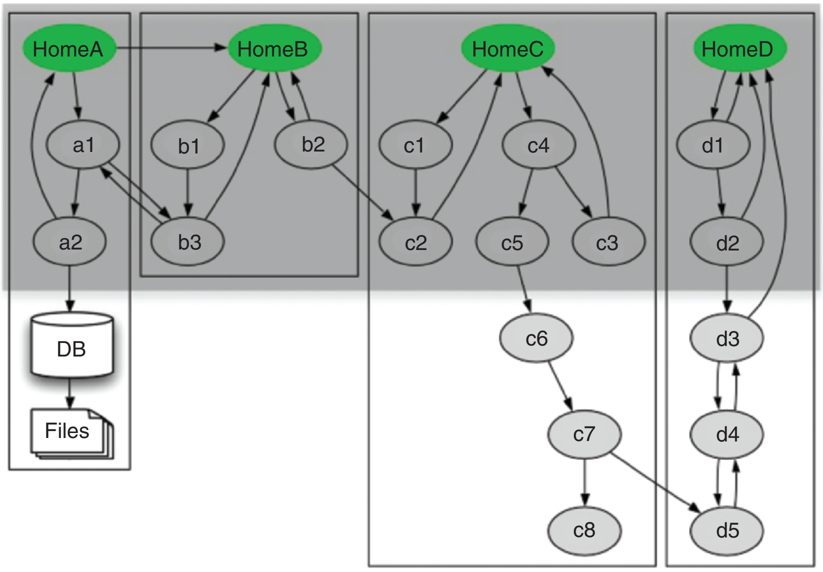


Fig. 1.8. Extensive collections, included more sites but archived at the surface level only. Only content in the shaded area will be archived. Pages deep in the hierarchy (c6, c7, c8, d3, d3, d5) as well as contend hidden behind database (hidden Web) will not be captured

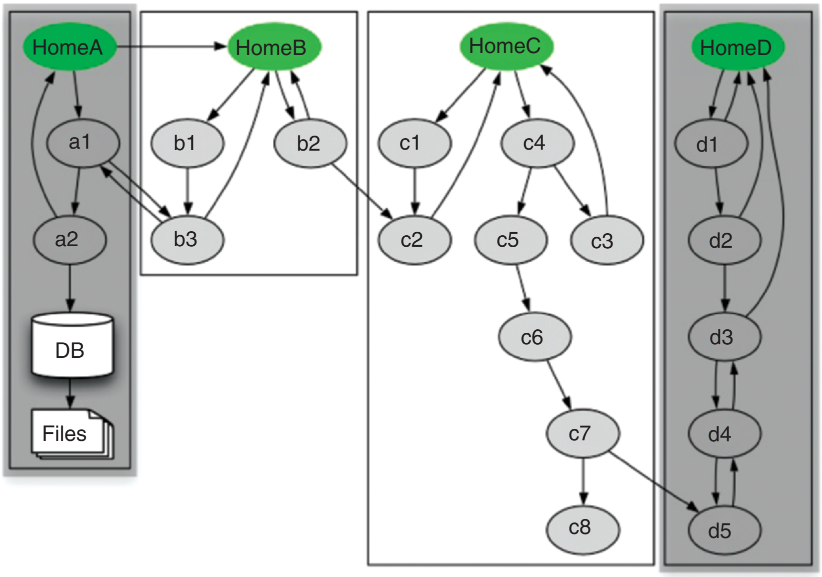


Fig. 1.9. Intensive archiving: less sites are crawled but crawl is done in depth. Only site A and D will be archived, but in totality, including the hidden Web portion of site A

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)