Web Archiving Methods and Approaches: A Comparative Study

Julien Masanès

Abstract

The Web is a virtually infinite information space, and archiving its entirety, all its aspects, is a utopia. The volume of information presents a challenge, but it is neither the only nor the most limiting factor given the continuous drop in storage device costs. Significant challenges lie in the management and technical issues of the location and collection of Web sites. As a consequence of this, archiving the Web is a task that no single institution can carry out alone. This article will present various approaches undertaken today by different institutions; it will discuss their focuses, strengths, and limits, as well as a model for appraisal and identifying potential complementary aspects amongst them. A comparison for discovery accuracy is presented between the snapshot approach done by the Internet Archive (IA) and the event-based collection done by the Bibliothèque Nationale de France (BNF) in 2002 for the presidential and parliamentary elections. The balanced conclusion of this comparison allows for identification of future direction for improvement of the former approach.

A Virtually Infinite Information Space

Assessing the size of the Web is a difficult task, and many attempts to provide a reliable estimate of it have been made so far with limited success. We will not review these attempts here but instead outline major changes the Web has introduced and discuss their impact for Web archiving.

Authorship Revolution

The Blog phenomenon is the most recent illustration of this revolution: the first Web browser designed and coded by Tim Berners Lee included

LIBRARY TRENDS, Vol. 54, No. 1, Summer 2005 ("Digital Preservation: Finding Balance," edited by Deborah Woodyard-Robinson), pp. 72–90

an authoring tool, which he considered to be an essential piece of the new system (Gillies & Caillau, 2000; Berners-Lee & Fischetti, 2000). Despite the subsequent omission of authoring tools from Web browsers, the Web has continued to offer an open publishing platform with global accessibility and continuous updating capacity.

This has dramatically changed the setting for publication, allowing almost anyone to bypass the traditional publishing actors and reach direct access to a potentially unlimited audience. The eventual impact of this change remains to be seen, but several consequences for archiving are already tangible.

The first important change is the end of an object's stability, with obvious impacts for archiving—an activity that in essence consists of capturing the state of an object at a point in time. The Web offers the ability to update content at any moment without notification (if additional notification mechanisms like Really Simple Syndication [RSS] protocol feeds are not in place), which poses a great challenge for archivists. Revisiting pages consumes resources, even if heuristics can be found to alleviate this process (Clausen, 2004). Choice of an appropriate frequency for capture can be problematic because, to be efficient, it should be done at the page level in most cases. It is indeed equivalent to assessing the probability of losing some intermediary updates between two captures.

Content Shaping

In addition to the change in the publication process, an important shift has occurred in the nature of documents themselves. The proliferation of citations that the hypertext environment allows induces a tremendous tendency toward dispersal of content, which archivists have to take into account in their approach. Web documents at the page level (but also the site level) hardly ever make sense alone. They are mingled in a larger document network that forms what Nelson named a "docuverse" (Nelson, 1992). From this perspective, archiving means extracting slices of the Web that constitute a whole metadocument (Landow, 1997); that is, spatially sampling the Web and making decisions each time regarding the exact perimeter of what to include, being aware that with time noninclusion means loss. For example, does archiving a site mean leaving out any document linked outside of its domain? If not, to what depth should external links be followed? There is no general answer to these questions, only specific ones based on the ultimate goal driving the archiving.

Choices also have to be made concerning what characteristics or functionalities are to be preserved. When a site is not primarily a collection of static pages, an archivist may focus on the interaction of functionalities (not only for navigation) and more generally the experience the site provides[1] in the archival context.

Convergence

It is worth noting that the Web is not only a platform absorbing previously existing Internet applications (mail, FTP, news) as well as non-Internet-based applications (database, document repository, and various information systems), but it also tends to be an entry point for almost everything today. This is a clear consequence of the design adopted for Uniform Resource Identifiers (URI), which Tim Berners-Lee insists is the most important standard of the Web (Berners-Lee & Fischetti, 2000). The prefix, the use of the Domain Name Server (DNS) system for host naming, and the flexibility offered for Webmasters regarding the right part of the URI, together make URI a powerful unifying standard. But for archivists this means almost everything can end up in their nets. If they want to focus on published material in the traditional sense of the word, they might want to filter online forums, for instance, or avoid diving into huge databases. Clues can be used for limiting the archiving, using, for instance, URI pattern detection (this has long been the case with search engines avoiding any dynamically generated content based on URI-embedded queries). This can extend to filtering content on the fly or during post-processing.

Technique

Even when the target is clearly identified and delimited, content acquisition can be an issue. Automatic tools for content gathering such as crawlers (also called spiders)[2] allow massive content acquisition at relatively low cost. With standard desktop computers and a Digital Subscriber Line (DSL) connection, it is possible today to retrieve millions of documents per week, even per day. Crawlers are also powerful and systematic tools for exploring the Web and discovering new sites through links even when starting from a very small set of seed sites.

There are severe crawler limitations, however, when it comes to finding a path to certain types of documents. First, access to sites or parts of sites can be restricted (with password or Internet Protocol [IP] authentication). In this case, getting authorization is needed. Second, the coding technique used to implement links can be hard to interpret for crawlers. This can be the case when scripts use contextual elements or when the code is opaque (executable, server-side code, etc.). Crawlers are getting better at link extraction[3] but still face some limits. Finally, a nontrivial interaction from the user can be required (that is, more than a click). This is usually the case when entering a query is required to access some portion of content.[4]

Content acquisition in this situation entails a case-by-case assessment, and adapted actions must be taken. This can be limited to entering new parameters for the crawler or downloading directly page by page some part of the site. In many cases still, nothing can be done remotely, and getting the content through the hypertext transfer protocol (http) interface is not possible. In these cases, pursuing direct contact with the producer

is unavoidable, which is extremely time consuming compared to direct online capture.

To summarize this quick overview of the situation,[5] we observe that the extraordinary extension of opportunity the Web offers for producers results in a corresponding increase in difficulty for archivists. Therefore, one should not be surprised to see that a variety of complementary approaches to Web archiving have been followed so far. The rest of this article proposes a comparative model of these approaches.

A Comparative Model

Approaches to Web archiving can be compared along several axes. Their scope, method, and level of quality can be different. Relative importance of manual and case-by-case handling compared to automatic and bulk processing of Web sites must also be considered.

Scope

Web archiving today is either site-, topic-, or domain-centric. Site-centric archiving is mostly done by corporate bodies, institutions, or even individuals for limited archiving purposes. We do not appraise this type of capture in this model as it does not entail collection building.

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. 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 or Archipol for analysis of Dutch political sites at Groningen University in the Netherlands. These projects share not only a topic orientation but also the use of a network of informants (Lecher, 2004); that is, researchers who provide accurate and updated feeds for the archive.

Other topic-centric projects have been carried out in libraries by actively seeking and archiving electoral Web sites, such as the Minerva project from the Library of Congress (Schneider, Foot, Kimpton, & Jones, 2003) or the French elections web archive fulfilled by the Bibliothèque nationale de France (BNF), which is discussed below. 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 separate activity.

Alternatively, domain-centric Web archiving is not driven by content but by content location. "Domain" is used here in the network sense of the word or, by extension, in the national sense of the term. Projects implementing this approach focus on a generic domain like .gov (Cruse , Eck-

man, & Kunze, 2003; Carlin 2004) or .edu (Lyle, 2004). It can also extend to a national domain, like Kulturarw started in 1997 by the Swedish Royal Library (Mannerheim, Arvidson, & Persson, 2000), which covers the .se domain and also Swedish pages linked from it and located in generic domains such as .com.

Methods

Projects can also noticeably differ with respect to the methodological approach they take for discovery, acquisition, and description of content. Automation of these tasks enables a tremendous lowering of the cost per site archived. Ideally, a single operator running a crawl can "discover" and download millions of sites through link detection and following. If we dare to assume that full-text indexing provides a powerful finding aid comparable if not superior to cataloguing, then we must conclude here again that automation lowers costs dramatically, as it can easily be applied on a large scale.

Unfortunately, automation reaches some limits, and manual handling must be done in certain 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 we saw with DACHS and Archipol. 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 (Bergmark, 2002; Bergmark, Lagoze, & Sbityakov, 2002), automatic tools can certainly not yet compare with a network of experts providing direct linking to the best material they are aware of.

However, a lack of "expertise" is not the only disadvantage crawlers have. Also to be considered is the delay needed to find new sites. The use of linking to discover new sites can be a long process in a global crawl. When a crawler comes to an ephemeral site, such as a site related to an event, for instance, the delay could be too long to locate and archive the related material.

This difference in efficiency between manual and automated discovery is, to our knowledge, undocumented in the literature. Later in this article we present elements for a comparison between sites discovered by Alexa's crawler and accessible today in the Internet Archive (IA) and sites related to the French elections of 2002 located by a team of reference librarians and archived at BNF.

Quality

The quality of a Web archive can be defined by (a) the completeness of material (linked files) archived within a designated perimeter and (b) being able to render the original form of the site, particularly regarding navigation and interaction with the user. Graphically, completeness can be measured horizontally by the number of relevant entry points found within

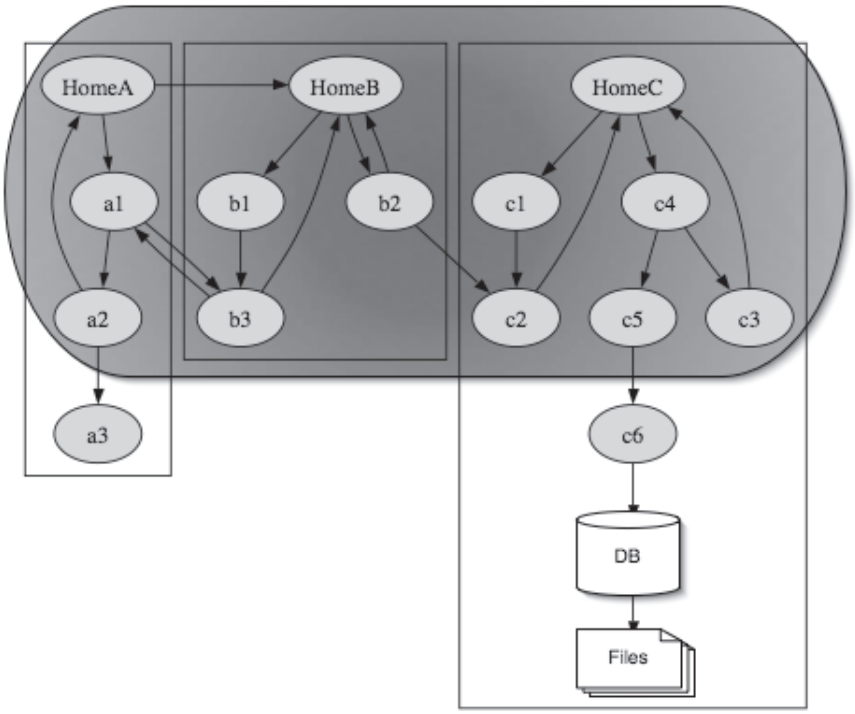

Figure 1. Extensive Archiving (Shaded Area). Some Pages Are Missed (a3, c6) as Well as the "Hidden" Part of Sites

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, verticality is limited to inline documents (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 nodes are not included based on their belonging to the site but only on their relevance to the topic.

Ideally, any archive should be complete vertically as well as horizontally. But, as we have seen, Web archiving is often a matter of choices, as perfect and complete archiving is unreachable. Archiving is called "extensive" when horizontal completeness is preferred to vertical completeness (see figure 1). This is the case, for instance, for the IA and its collection, which is donated by Alexa (as Burner [1997] explains, Alexa's crawler uses a breadth-first approach and adapts depth of crawl for a site according to traffic measured for this site).

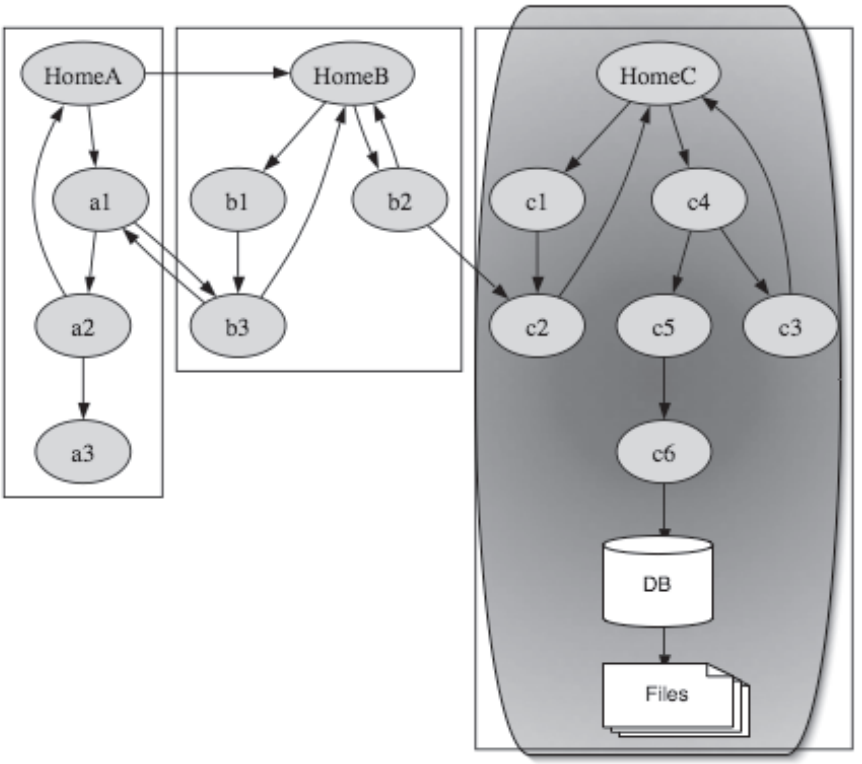

Figure 2. Intensive Archiving (Shaded Area). Aims to Collect Fewer Sites But Collects Deeper Content, Including Potentially "Hidden" Web

Conversely, archiving is called "intensive" when vertical completeness is preferred to horizontal completeness (see figure 2). This is the case, for instance, when a site-first priority is used for crawlers (Masanès, 2004) or when a manual verification with supplementary archiving is made where needed. 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 (see some experiment in this area in Masanès 2002a, 2005a).

It should be noted that there are ways of escaping from a purely binary choice (intensive vs. extensive). For instance, crawler accuracy has been measured to appraise the best balance between depth of crawls and coverage. Baeza-Yates and Castillo (2004) have recently shown that crawling five levels deep is enough to reach 90 percent of the useful content in a Web site. This kind of estimate provides a larger range of choices. It has also been argued by Masanès (2002b) that a temporal combination of policy

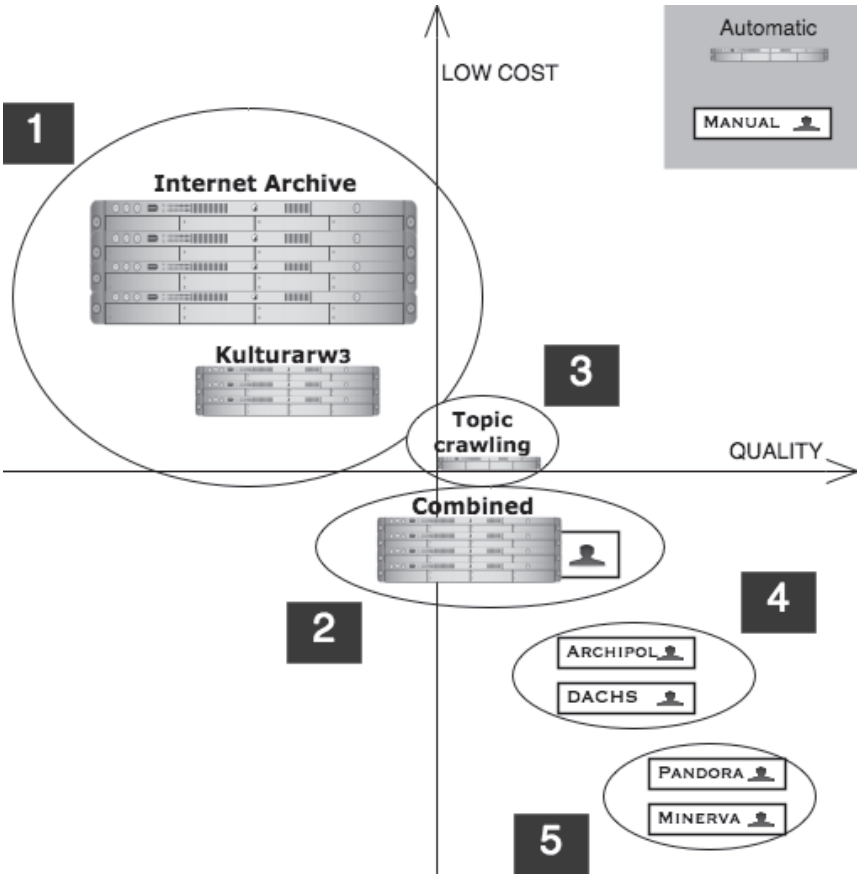

Figure 3. A Cost/Quality Comparison of Web Archiving Approaches

(extensive twice a year and focused intensive in between) can provide better overall results even when done entirely automatically.

Different Approaches to Web Archiving

From the previous section, we can draw a comparison between various documented Web archiving projects.[6] It should be clearly stated that the aim of this comparison is not to judge them but to provide a better understanding of the diversity in methodological approaches to Web archiving as well as potential complementary aspects between them.

Cost/Quality Comparison

The first comparison presents the economic positioning of these approaches through a traditional cost/quality comparison. Figure 3 shows

several clusters with examples of projects for each. Crawling-based projects (for discovery) are represented by a server pictogram, and manual-based approaches are represented by a small portrait within a label. On the x-axis projects increase by quality, and on the y-axis projects increase by cost.

Cluster 1 is characterized by a very low cost per site archived but quite low quality. Cluster 1 includes a large domain-centric snapshot like Kulturarw3 from the Swedish National Library and the even larger IA snapshots. Processing is entirely automated and very little if any quality verification is done except crawl monitoring. On the other hand, they provide very wide coverage of their designated domain. Cluster 2 groups combined approaches also applied at the domain level. They aim to take advantage of crawling with complementary acquisition of content. For example, BNF's approach includes "hidden" Web site deposit. While this improves the completeness and hence the quality of the archive, it also significantly increases costs. Cluster 3 includes topic crawling-oriented projects, which combine quite low cost with reasonable quality as they often adopt a purely horizontal perspective. Cluster 4 assembles topic-centric projects done manually but based on informant networks that allow them to be less expensive than other manually handled projects. This is the case for Archipol as well as for DACHS. However, as they are undertaken in smaller and less specialized structures, we can make the hypothesis that they provide a lesser degree of verification and overall quality, but this remains to be demonstrated. Finally, Cluster 5 groups domain- or topic-centric projects undertaken by libraries without informant networks but with dedicated staff that provide a manual verification of archived sites. This implies a higher cost per site than other approaches while providing a better overall quality of site archived. In the diagram all these projects fit along a diagonal and, of course, the direction for improvement is clearly toward the upper-right-hand corner.

Collection-Oriented Comparison

The second comparison we propose (see figure 4) is based on collection orientation. The x-axis depicts the orientation of the collection regarding preferred completeness as defined earlier.

On the left side of the graph, archiving is made in extension (horizontal completeness preferred) and on the right in intension (vertical completeness preferred). The y-axis shows the orientation of the collection with regard to its target (domain or topic). Upper clusters (1, 2, and 3) represent the domain-centric approaches taken by national libraries. They also include a domain-centric crawl of the .gov sites done by the National Archives and Records Administration (NARA) in the United States before the new presidential electoral mandate (Carlin, 2004). Topic-centric clusters (4, 5, and 6) are on the bottom of the figure.

Extensive projects are located on the left side of the figure, intensive ones on the right. It is apparent that the intensive approach goes with a

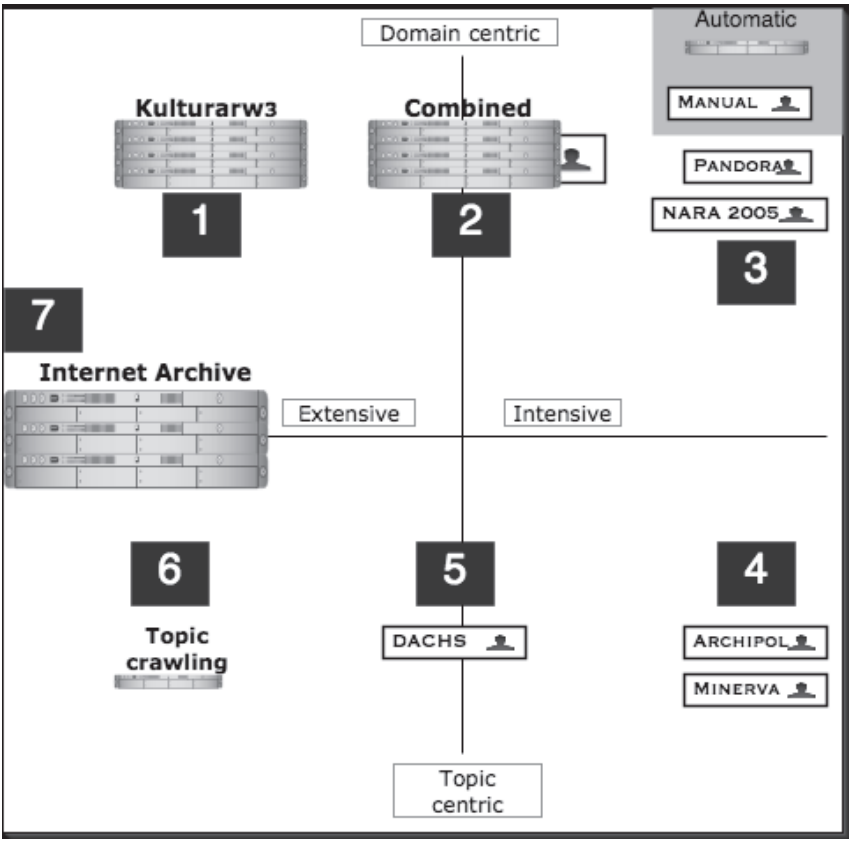

Figure 4. Collection Orientation of Web Archiving Approaches

manual discovery of URI seeds. Starting from a limited set of entry points, these projects tend to focus on depth. Most of them, however, do not touch the Deep Web where manual intervention cannot be replaced by any other means while, for discovery, manual selection is increasingly challenged by automated tools. Could budgetary allowances for manual discovery be better employed for breaking the Deep Web limit? Experience in this domain lacks a clear response.

In the following sections we present some results that show where manual selection can add value in the context of topic archiving by comparing manual selection and the snapshot method for event-related collections. As mentioned above, it is expected that domain snapshots will provide an insufficient coverage of very ephemeral sites; for example, those that appear in relation to an event. This is the case for many political campaign Web sites, whose lifespans may only be a few months or weeks. If this proves

to be the case, manual and active discovery of sites would be justified to build event-related collections even to complement a domain snapshot as undertaken by national libraries.

The following section presents results to assess precisely the added value of manual discovery for event-based collections. We compared the collection of the IA as provided by Alexa and the collection the BNF built in 2002 for the presidential and parliamentary elections. The method of building the two collections was very different. The Internet Archive collection is built from two-month snapshots done by Alexa for the entire Web.[7] The total size of these snapshots ranges up to tens of terabytes each.

BNF's collections were selected manually and information was stored in a database comprising the Web sites' URIs, their archiving frequency, and several other fields aimed mostly at organizing the work at the BNF (like candidate, party, location, and type of site). The acquisition of sites was done using a simple tool based on HTTrack.[8]

Obviously this method will not scale up for a large domain archive (Phillips, 2003), but it was initially used by BNF because we were not able to do large-scale domain crawling at the time. It was also expected that a large national domain crawl would not efficiently capture the sites at stake within the available time scale. This latter assumption was correct as the following results show.

Methodology

BNF's team selected 696 entry points for the presidential elections and 1,002 for the parliamentary elections. This does not include entry points that were selected for both elections (including permanent party Web sites, for instance, considered to be not event-related in this context). Table 1 shows the distribution of entry points between sites and parts of sites (sections).

A script[9] was used to test the match between these entry points and the IA's Wayback Machine collections for three different periods: 2001, 2003, and three months of 2002 (March to May 2002 for the presidential elections, May to July for the parliamentary elections). This allows us to determine (a) if the entry point already existed in 2001, which can be deduced from its presence in IA for this year; (b) if IA had at least one version of the entry point for a time near the elections; and (c) if its URI disappeared or was found later (in 2003) by Alexa's crawler.

It should be noted that this protocol only gives us an indication of the relative accuracy of the two methods (manual discovery and domain snapshot) given that we only tested the IA collection for the presence of at least one version of the entry point for the entire period of the elections (and generally, when there was one, it was usually the only one). For many of the sites, however, the BNF made several copies (up to weekly) during the three months of the campaign. We have not compared the vertical completeness of the two methods. Neither do we compare how the Websphere

Table 1. Distribution of Entry Points between Sites and Parts of Sites or Sections

	Sites	Parts of Sites or Single Documents (Sections)	Total Entry Points
Presidential Elections Only	182	514	696
Parliamentary Elections Only	604	398	1002

of these sites was archived,[10] given that the BNF made no attempt to follow links horizontally and therefore only achieved accidental coverage of the Websphere of these sites.

Results

The following items were calculated:

- Number of entry points (EP) present in IA's collection for 2001, 2002 (campaign), and 2003; this gives an indication of the coverage achieved by the crawler compared to what had been manually selected
- Number of EP discovered in 2002 by the crawler (not present in 2001 and present in 2002 during the campaign)
- Ratio of EP discovered in 2002 compared to nonexisting EP in 2001; this gives a good indication of the capacity of the crawler to timely discover items; that is, when they appear during the campaign
- EP lost in 2002 (present in 2001 and not in 2002 during the campaign); this gives an indication of the crawling process' irregularity and erratic behavior
- Balance between discovered and lost EP (relative to the nonexisting or to be discovered EP in 2001); this global accuracy measurement is the final outcome of this comparison: it takes into account the ability of the crawling process to discover EP in a timely manner, balanced by the measured irregularity it has shown in this context

Figure 5 shows IA's coverage for EP of the presidential election. The results clearly indicate a good coverage for EP at the site level (in column 1 of 3). This coverage shows a linear improvement from 2001 to 2003 and reaches two-thirds during the campaign.

The situation at the section level (column 2 of 3) is more erratic. Coverage is worse than it is at the site level (15.4 vs. 50; 4.3 vs. 67; and 51.9 vs. 83) but its evolution is even more noticeable: the coverage falls by more than three times between 2001 and 2002. This means that, even if 79 EP were present in IA out of 514 in 2001, this number falls to 22 in 2002. As 14 were discovered in the same time, this means that 71 were lost from 2001 to 2002. This is not due to a weakness in discovery (even if it is very weak in this case) as we do not observe such an evolution at the site level. We must conclude that this is related to the vertical completeness of the crawl, which seems to vary a lot from one snapshot to another.

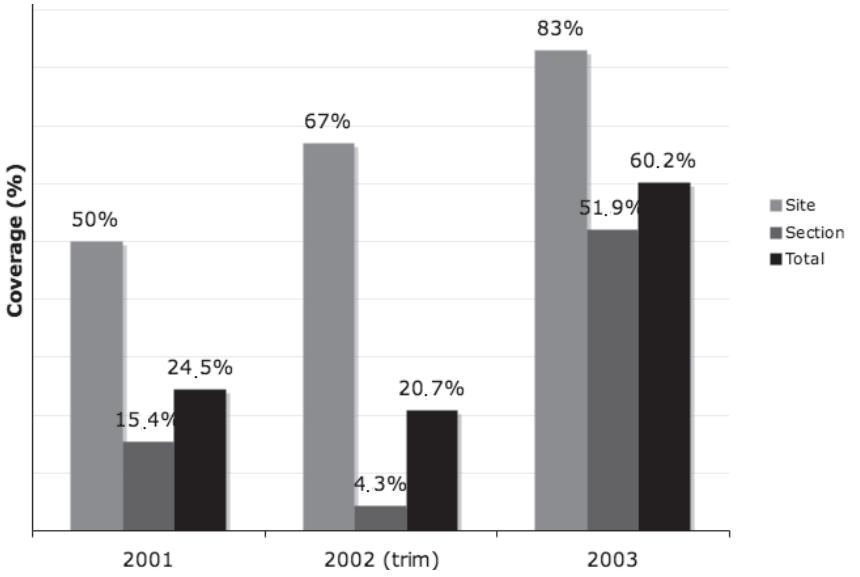

Figure 5. IA's Coverage compared to BNF's selection for the 2002 French Presidential Election

These observations are confirmed by the related entry points for the parliamentary elections, which show the same pattern of evolution at the site and section level, although the results are better at the section level in this case (see figure 6). This coverage is considered satisfactory at the site level. For the presidential elections, where only sixteen candidates were competing, sites selected were mostly secondary sites (analysis, comments, humoristic, critics, etc.). For the parliamentary elections, however, where thousands of candidates were competing, we had to focus our selection on candidate Web sites. In both cases, at the site level more than two-thirds coverage is achieved, which is quite acceptable.

However, when it comes to more precise selection (sections of sites or even single documents) the results are in favor of manual selection. There is no reason to believe that discovery of the root level of these EP is more difficult, so we also measure the vertical completeness of Alexa's crawl, which seems to have been limited and taken longer to achieve (51.9 percent for the presidential elections and 52 percent for the parliamentary elections for coverage achieved in 2003 only).

For measuring the crawler's ability (when used for a domain crawl) to discover EP in a particular period, we have excluded from our set EP that already existed in IA in 2001 and measured the proportion of the remaining ones that were discovered during the campaign (2002). The resulting "discovery accuracy" is shown in figure 7.

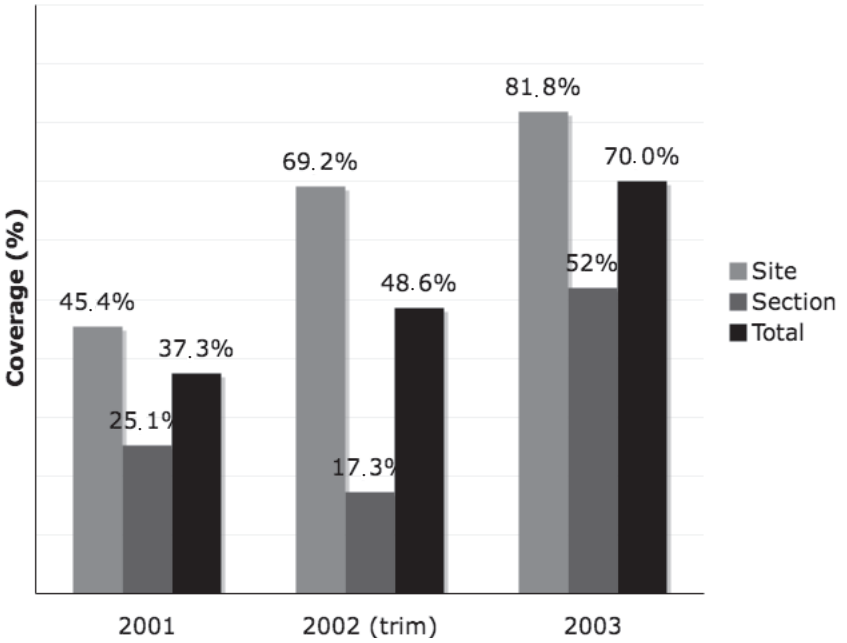

Figure 6. IA's Coverage Compared to BNF's Selection for the 2002 French Parliamentary Election

We see that discovery accuracy is consistent for both elections despite the difference in composition of each set mentioned above (one comprises many secondary sites, the other mostly candidate sites). It also shows a clear distinction between the site level (approaching two-thirds) and the section level, which is much lower. This is consistent with the results for overall coverage discussed above. This enhances the relatively good result obtained at the site level by showing that the discovery of unknown sites is done within the three months of the campaign in almost two-thirds of the cases. At the section level, however, the results drop dramatically to less than one-fifth.

However, these results do not accurately reflect the comparison of the crawler to manual selection because they do not take into account the loss of EP between 2001 and 2002. To balance this we have calculated the global accuracy (the discovery of EP minus the loss of EP divided by the number of EP to be discovered in 2001). Figure 8 shows that these results are significantly less favorable for the crawler.

The results at the site level drop noticeably (61.5 to 34.1 and 64.8 to 43.6), which is due to a high level of loss, even at the site level, between 2001 and 2002. This is even more obvious at the section level, where the final balance between discovery and loss is negative for both elections (–13.1 percent and –10.4 percent). This result highlights the already mentioned

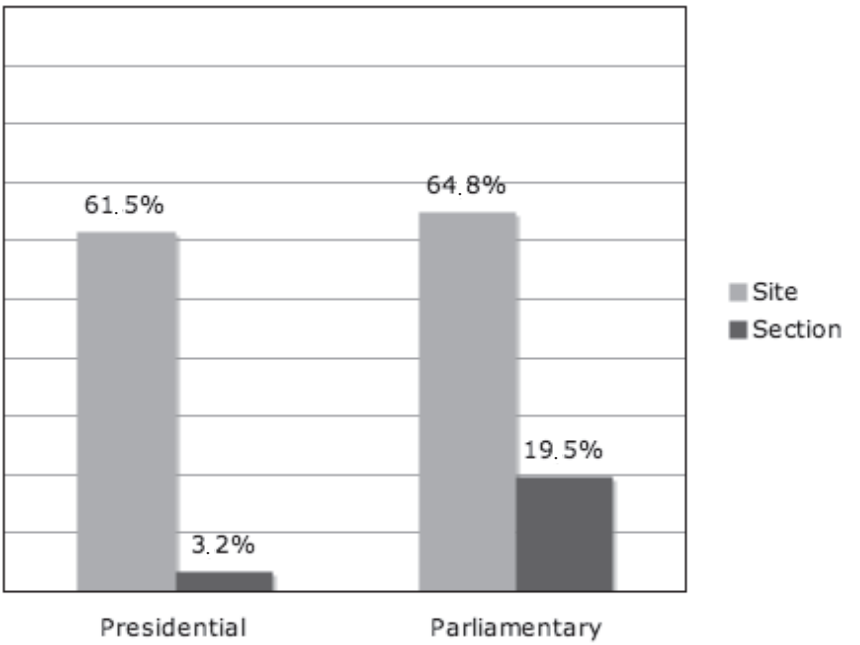

Figure 7. Discovery Accuracy: A Measurement of the Crawler's Capability for Timely Discovery of Sites and Sections of Sites

irregularity of crawlers during large-scale snapshots. We do not know if Alexa uses an extensive list of EP already crawled to start a new crawl and, even if these results tend to indicate that this is not the case, it could also be a consequence of lack of time to revisit already crawled sites. However, this introduces a serious restriction in the positive results we found for pure discovery. If crawlers tend to find new EP successfully, they also show a strong irregularity that can result in loss of EP.

Conclusion

We have seen that the challenges associated with Web archiving require consideration and appraisal of a variety of approaches that can complement each other and allow better global efficiency for preservation of Web content. This implies definition of key parameters for an archiving policy (such as prefered completeness, overall quality, cost per site, orientation). It also suggests how to measure discovery, acquisition, verification, and preservation of content with relative accuracy. We present initial results for the former and show that, during a snapshot crawl, ephemeral sites tend to be discovered with relative accuracy as long as the temporal window is sufficently large (three months in this case).

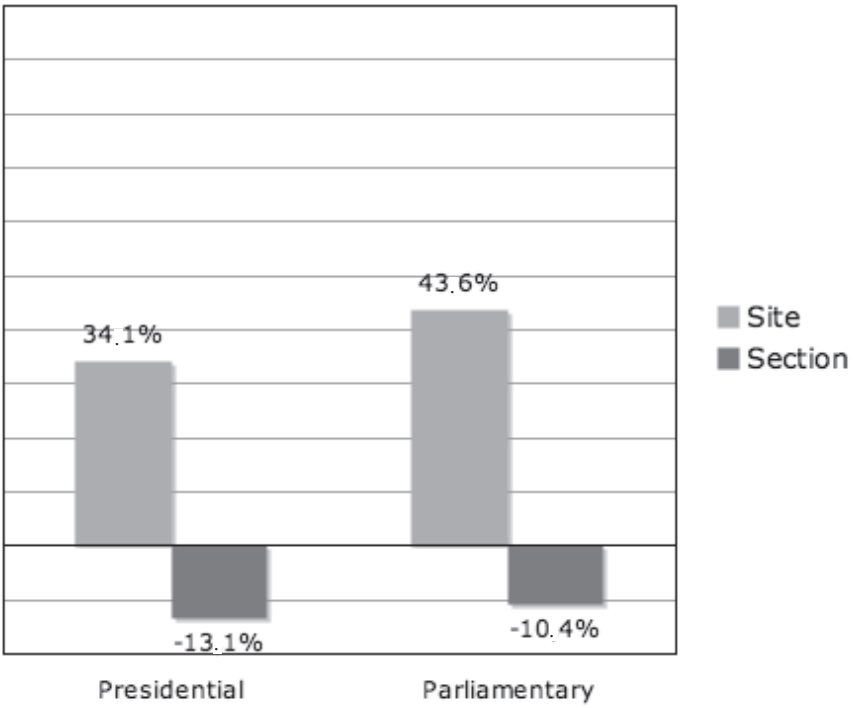

Figure 8. Global Accuracy of Crawler for Sites/Sections of Sites Discovery

But these results have to be balanced by the fact that crawlers tend to present strong irregularity in their horizontal coverage. This includes loss of EP when they start from a "fresh" list of seed URIs, which is often the case in the search engine world for sake of adaptation to current Web topology. It can be argued, however, that Web archiving crawlers could be used in a more conservative manner, ensuring the continuity of coverage as a priority. The second balance to take into account is the lack of depth that large-scale snapshots tend to present. This is particularly noticeable when it comes to selection of sections of sites related to an event. This demonstrated a distinct weakness of the crawlers coverage compared to manual selection.

Finally, we should emphasize that event-related Web sites present a higher frequency of change (this is the case also for news-related Web sites) and therefore need to be archived more often. Manual selection often includes estimation of an appropriate archiving frequency, which is then implemented by crawlers. Such comparisons should be updated over time as crawler functionality and use scenarios evolve. We hope this work will contribute to laying the ground for Web archiving appraisal and adaptation in the future.

Notes

1. For a general presentation of the different angles the Web can be considered from, see Burnett & Marshall (2003). For the Internet considered as an object of study from a noncontent perspective, see Hine (2000). For the Internet considered from the archival perspective, see Christensen-Dalsgaard (2001).
2. For a recent overview of crawling technology, see Pant, Srinivasan, & Menczer (2003). See also Chakrabarti (2002).
3. See, for instance, Heritrix, the International Internet Preservation Consortium (IIPC) official crawler developed jointly by Internet Archive and the several Nordic national Libraries (Mohr, Kimpton, Stack, & Ranitovic, 2004).
4. See the IIPC (netpreserve.org) surveys on this issue (Boyko 2004). See also Marill, Boyko, & Ashenfelder (2004).
5. For a more detailed presentation on this see Masanès (2005b).
6. The following projects will be used as illustration. Several others not mentionned here are documented (see, for instance, the International Web Archiving Workshop series http://iwaw.net).

   Archipol (http://www.archipol.nl/) by the Dutch Documentatiecentrum Nederlandse Politieke Partijen (DNPP) (see Voerman, Keyzer, Hollander, & Druiven, 2002)
   Bibliothèque nationale de France has adopted a combined approach including archiving of Deep Web sites (see Masanès, 2002a, 2002b; and Abiteboul, Cobena, Masanès, & Sedrati, 2002; see also the Danish combined approach as explained by Christensen-Dalsgaard, 2004).
   DACHS (see Lecher, 2004; Gross, 2003)
   Internet Archive (http://www.archive.org) (see Kahle, 1997; Burner, 1997)
   Kulturarw3 (http://www.kb.se/kw3/ENG/Default.htm) by the National Library of Sweden (see Mannerheim et al., 2000; Arvidson, 2002)
   Minerva (http://www.loc.gov/minerva/) by the Library of Congress (see Arms, Adkins, Ammen, & Hayes, 2001; Schneider, Foot, Kimpton, & Jones, 2003)
   NARA, governmental agencies Web site archiving, 2004–2005 (see Carlin, 2004)
   PANDORA (http://pandora.nla.gov.au) by the National Library of Australia (see Phillips, 2003; Koerbin, 2004)
   Topic Crawling (see Bergmark, 2002)

7. For an appraisal of national biases of IA's collection and a discussion of their origin, see Thelwall & Vaughan (2004).
8. See http://www.httrack.com/ for more information.
9. Thanks to Younès Hafri from BNF for handling this.
10. See Foot and Schneider (2002) for a definition of the concept of Websphere and its application for election Web sites in the U.S. campaign.

References

Abiteboul, S., Cobena, G., Masanès, J., & Sedrati, G. (2002). *A first experience in archiving the French Web.* Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries, Rome, Italy, September 16–18.

Arms, W. Y., Adkins, R., Ammen, C., & Hayes, A. (2001). Collecting and preserving the Web: The Minerva prototype. *RLG DigiNews, 5*(2). Retrieved March 16, 2005, from http://www.rlg.org/preserv/diginews/diginews5-2.html.

Arvidson, A. (2002). *The collection of Swedish Web pages at the Royal Library—The Web heritage of Sweden.* Paper presented at the 68th IFLA Council and General Conference, Glasgow, UK, August 18–24.

Baeza-Yates, R. A., & Castillo, C. (2004). Crawling the infinite Web: Five levels are enough. In *Workshop on algorithms and models for the Web-Graph WAW2004.* Rome, Italy: Springer Verlag.

Bergmark, D. (2002). *Collection synthesis.* Paper presented at the 2nd ACM/IEEE-CS Joint Conference on Digital libraries, Portland, July 14–18.

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, Rome, Italy, September 16–18.

Berners-Lee, T., & Fischetti, M. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor.* New York: HarperCollins.
Boyko, A. (2004). Test bed taxonomy. In IIPC (Ed.), *IIPC Reports.* Paris: IIPC.
Burner, M. (1997). Crawling towards eternity: Building an archive of the World Wide Web. *Web Techniques Magazine, 2*(5). Retrieved March 16, 2005, from http://www.webtechniques.com/archives/1997/05/burner.
Burnett, R., & Marshall, P. D. (2003). *Web theory: An introduction.* New York: Routledge.
Carlin, J. W. (2004). Harvest of agency public Web sites. *NARA Bulletin, 2005–02.* Retrieved March 16, 2005, from http://www.archives.gov/records_management/policy_and_guidance/bulletin_2005_02.html.
Chakrabarti, S. (2002). *Mining the Web: Discovering knowledge from hypertext data.* San Francisco, CA: Morgan Kaufmann.
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, June 18–19.
———. (2004). Web Archive activities in Denmark. *RLG DigiNews, 8*(3). Retrieved March 16, 2005, from http://www.rlg.org/en/page.php?Page_ID=17661.
Clausen, L. (2004). *Concerning etags and datestamps.* Paper presented at the 4th International Web Archiving Workshop, Bath (UK), September 16.
Cruse, P., Eckman, C., & Kunze, J. (2003). *Web-based government information: Evaluating solutions for capture, curation, and preservation.* Oakland: California Digital Library.
Foot, K., & Schneider, S. M. (2002). Online Action: In the US 2000 political campaign. In V. Serfaty (Ed.), *L'internet en politique des Etats-Unis à l'Europe* (p. 424). Strasbourg, France: Presses Universitaires de Strasbourg.
Gillies, J., & Caillau, R. (2000). *How the Web was born: The story of the World Wide Web.* Oxford: Oxford University Press.
Gross, J. (2003). *Learning by doing: The Digital Archive for Chinese Studies (DACHS).* Paper presented at the 3rd Workshop on Web Archives, Trondheim, Norway, August 21.
Hine, C. (2000). *Virtual ethnography.* Thousand Oaks, CA: Sage.
Kahle, B. (1997). Preserving the Internet. *Scientific American, 397,* 82–84.
Koerbin, P. (2004). *The PANDORA Digital Archiving System (PANDAS) and managing Web archiving in Australia: A case study.* Paper presented at the 4th International Web Archiving Workshop, Bath (UK), September 16.
Landow, G. P. (1997). *Hypertext 2.0* (rev. ed.). Baltimore: Johns Hopkins University Press.
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 Institutes, NLA, Canberra, Australia.
Lyle, J. A. (2004). *Sampling the Umich.edu domain.* Paper presented at the 4th International Web Archiving Workshop, Bath (UK), September 16.
Mannerheim, J., Arvidson, A., & Persson, K. (2000). *The Kulturarw3 project—The Royal Swedish Web Archiw3e—An example of "complete" collection of Web pages.* Paper presented at the 66th International Federation of Library Associations and Institutions Council and General Conference, Jerusalem, Israel, August 13–18.
Marill, J., Boyko, A., & Ashenfelder, M. (2004). *Web Harvesting Survey, v. 1.* Retrieved March 16, 2005, from http://www.netpreserve.org/publications/iipc-r-001.pdf.
Masanès, J. (2002a). *Archiving the deep Web.* Paper presented at the 2nd International Workshop on Web Archives, Rome, Italy, September 19.
———. (2002b). Towards continuous Web archiving: First results and an agenda for the future. *D-Lib Magazine, 8*(12). Retrieved March 16, 2005, from http://www.dlib.org/dlib/december02/masanes/12masanes.html.
———. (2004). *Site-first priority: Implementing the frontline.* Paris: IIPC.
———. (2005a). Collecting the hidden Web. In J. Masanès (Ed.), *Web Archiving.* New York: Springer Verlag.
——— (Ed.). (2005b). *Web Archiving.* New York: Springer Verlag.
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, Bath (UK), September 16.
Nelson, T. H. (1992). *Literary machines: The report on, and of, Project Xanadu concerning word process-*

ing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom (ed. 93.1). Sausalito, CA: Mindful Press.

Pant, G., Srinivasan, P., & Menczer, F. (2003). Crawling the Web. In M. Levene & A. Poulovassilis (Eds.), *Web Dynamics* (pp. 153–78). New York: Springer-Verlag.

Phillips, M. E. (2003). Collecting Australian online publications. *Australian Academic and Research Libraries (AARL), 34*(3). Retrieved August 15, 2005, from http://alia.org.au/publishing/arl/34.3/full.text/phillips.html.

Schneider, S. M., Foot, K., Kimpton, M., & Jones, G. (2003). *Building thematic Web collections: Challenges and experiences from the September 11 Web archive and the election 2002 Web archive.* Paper presented at the 3rd ECDL Workshop on Web Archives, Trondheim, Norway, August 21.

Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet archive. *Library & Information Science Research, 26*(2), 162–76.

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). Retrieved March 16, 2005, from http://www.essex.ac.uk/ecpr/publications/eps/onlineissues/autumn2002/information.htm.

Julien Masanès, IIPC Programme Officer, Bibliothèque nationale de France, quai François Mauriac, 75706 Paris cedex 13, France, julien.masanes@netpreserve.org. Julien Masanès studied Philosophy and Cognitive Science, gaining his M.S. in Philosophy from the Sorbonne in 1992 and his M.S. in Cognitive Science from the Ecole des Hautes Etudes en Sciences Sociales (EHESS) in 1994. Mr. Masanès commenced work at the Bibliothèque Nationale de France (BNF) in 1999 in the area of Web archiving and obtained his M.S. in Librarianship from the Ecole Nationale Supérieure des Sciences de l'information et des Bibliothèques (ENSSIB) in 2000. At the BNF, Mr. Masanès organized the European Conference on Digital Libraries (ECDL) workshops on Web archiving (IWAW) and participated in the Networked European Deposit Library (NEDLIB) project. He is currently the conservator in the BNF digital library department, where he is responsible for the expansion of legal deposit to encompass Internet resources. He actively participated in the creation of the International Internet Preservation Consortium (IIPC), which he now coordinates.