A First Experience in Archiving the French Web

S. Abiteboul[1,3], G. Cobéna[1], J. Masanes[2], and G. Sedrati[3]

[1] INRIA: Serge.Abiteboul@inria.fr,
Gregory.Cobena@inria.fr
[2] BnF: Julien.Masanes@bnf.fr
[3] Xyleme: Gerald.Sedrati@xyleme.com

Abstract. The web is a more and more valuable source of information and organizations are involved in archiving (portions of) it for various purposes, e.g., the Internet Archive *www.archive.org*. A new mission of the French National Library (BnF) is the "dépôt légal" (legal deposit) of the French web. We describe here some preliminary work on the topic conducted by BnF and INRIA. In particular, we consider the acquisition of the web archive. Issues are the definition of the perimeter of the French web and the choice of pages to read once or more times (to take changes into account). When several copies of the same page are kept, this leads to versioning issues that we briefly consider. Finally, we mention some first experiments.

1 Introduction

Since 1537[4], for every book edited in France, an original copy is sent to the Bibliothèque nationale de France (French National Library - BnF in short) in a process called *dépôt légal*. The BnF stores all these items and makes them available for future generations of researchers. As publication on the web increases, the BnF proposes providing a similar service for the French web, a more and more important and valuable source of information. In this paper, we study technical issues raised by the legal deposit of the French web. The main differences between the existing legal deposit and that of the web are the following:

1. the number of content providers: On the web, anyone can publish documents. One should compare, for instance, the 148.000 web sites in ".fr" (as of 2001) with the 5000 traditional publishers at the same date.
2. the quantity of information: Primarily because of the simplicity of publishing on the web, the size of content published on the French web is orders of magnitude larger than that of the existing legal deposit and with the popularity of the web, this will be more and more the case.
3. the quality: Lots of information on the web is not meaningful.
4. the relationship with the editors: With legal deposit, it is accepted (indeed enforced by law) that the editors "push" their publication to the legal deposit. This "push" model is not necessary on the web, where national libraries can themselves find relevant information to archive. Moreover, with the relative freedom of publication, a strictly push model is not applicable.

[4] This was a decision of King François the 1st.

5. updates: Editors send their new versions to the legal deposit (again in push mode), so it is their responsibility to decide when a new version occurs. On the web, changes typically occur continuously and it is not expected that web-masters will, in general, warn the legal deposit of new releases.
6. perimeter: The perimeter of the classical legal deposit is reasonably simple, roughly *the contents published in France*. Such notion of border is more delusive on the web.

For these reasons, the legal deposit of the French web should not only rely on editors "pushing" information to BnF. It should also involve (because of the volume of information) on complementing the work of librarians with automatic processing.

There are other aspects in the archiving of the web that will not be considered here. For instance, the archiving of sound and video leads to issues of streaming. Also, the physical and logical storage of large amounts of data brings issues of long term preservation. How can we guarantee that terabytes of data stored today on some storage device in some format will still be readable in 2050? Another interesting aspect is to determine which services (such as indexing and querying) should be offered to users interested in analyzing archived web content. In the present paper, we will focus on the issue of obtaining the necessary information to properly archive the web.

The paper describes preliminary works and experiments conducted by BnF and INRIA. The focus is on the construction of the web archive. This leads us to considering issues such as the definition of the perimeter of the French web and the choice of pages to read one or more times (to take changes into account). When several copies of the same page are kept, this also leads to versioning issues that we briefly consider. Finally, we mention some first experiments performed with data provided by Xyleme's crawls of the web (of close to a billion pages).

In Section 2, we detail the problem and mention existing work on similar topics. In Section 3, we consider the building of the web archive. Section 4 deals with the importance of pages and sites that turn out to play an important role in our approach. In Section 5, we discuss change representation, that is we define a notion of delta per web site that we use for efficient and consistent refresh of the warehouse. Finally we briefly present results of experiments.

2 Web Archiving

The web keeps growing at an incredible rate. We often have the feeling that it accumulates new information without any garbage collection and one may ask if the web is not self-archiving? Indeed, some sites provide access to selective archives. On the other hand, valuable information disappears very quickly as community and personal web pages are removed. Also the fact that there is no control of changes in "pseudo" archives is rather critical, because this leaves room for revision of history. This is why several projects aim at archiving the web. We present some of them in this section.

2.1 Goal and scope

The web archive intends providing future generations with a representative archive of the cultural production (in a wide sense) of a particular period of Internet history. It may be used not only to refer to well known pieces of work (for instance scientific articles) but also to provide material for cultural, political, sociological studies, and even to provide material for studying the web itself (technical or graphical evolution of sites for instance). The mission of national libraries is to archive a wide range of material because nobody knows what will be of interest for future research. This also applies to the web. But for the web, exhaustiveness, which is required for traditional publications (books, newspapers, magazines, audio CD, video, CDROM), can't be achieved. In fact, in traditional publication, publishers are actually filtering contents and an exhaustive storage is made by national libraries from this filtered material. On the web, publishing is almost free of charge, more people are able to publish and no filtering is made by the publishing apparatus. So the issue of selection comes again but it has to be considered in the light of the mission of national libraries, which is to provide future generations with a large and representative part of the cultural production of an era.

2.2 Similar projects

Up to now, two main approaches have been followed by national libraries regarding web archiving. The first one is to select manually a few hundred sites and choose a frequency of archiving. This approach has been taken by Australia [15] and Canada [11] for instance since 1996. A selection policy has been defined focusing on institutional and national publication.

The second approach is an automatic one. It has been chosen by Nordic countries [2] (Sweden, Finland, Norway). The use of robot crawler makes it possible to archive a much wider range of sites, a significant part of the surface web in fact (maybe 1/3 of the surface web for a country). No selection is made. Each page that is reachable from the portion of the web we know of will be harvested and archived by the robot. The crawling and indexing times are quite long and in the meantime, pages are not updated. For instance, a global snapshot of the complete national web (including national and generic domain located sites) is made twice a year by the royal library of Sweden. The two main problems with this model are: (i) the lack of updates of archived pages between two snapshots, (ii) the deep or invisible web [17, 3] that can't be harvested on line.

2.3 Orientation of this experiment

Considering the large amount of content available on the web, the BnF deems that using automatic content gathering method is necessary. But robots have to be adapted to provide a continuous archiving facility. That is why we have submitted a framework [13] that allows to focus either the crawl or the archiving, or both, on a specific subset of sites chosen in an automatic way. The robot is driven by parameters that are calculated on the fly, automatically and at a large scale. This allows us to allocate in an optimal manner the resources to crawling and archiving. The goal is twofold: (i) to cover a very large portion of the French web (perhaps "all", although all is an unreachable notion because of dynamic pages) and (ii) to have frequent versions of the sites, at least for a large number of sites, the most "important" ones.

It is quite difficult to capture the notion of importance of a site. An analogy taken from traditional publishing could be the number of in-going links to a site, which makes it a publicly-recognized resource by the rest of the web community. Links can be consider similar, to a certain extent of course, to bibliographical references. At least they give a web visibility to documents or sites, by increasing the probability of accessing to them (cf the random surfer in [5]). We believe that it is a good analogy of the public character of traditionally published material (as opposed to unpublished, private material for instance) and a good candidate to help driving the crawling and/or archiving process [13]. Some search engines already use importance to rank query results (like Google or Voila).

These techniques have to be adapted to our context, that is quite different. For instance, as we shall see, we have to move from a page-based notion of importance to a site-based one to build a coherent Web archive. (see Section 4). This also leads to exploring ways of storing and accessing temporal changes on sites (see Section 5) as we will no longer have the discrete, snapshot-type of archive but a more continuous one. To explore these difficult technical issues, a collaboration between BnF and INRIA started last year. The first results of this collaboration are presented here. Xyleme provided different sets of data needed to validate some hypothesis, using the Xyleme crawler developed jointly with INRIA. Other related issues, like the deposit and archiving of sites that can not be harvested online will not be addressed in this paper [12].

One difference between BnF's legal deposit and other archive projects is that it focuses on the French web. To conclude this section, we consider how this simple fact changes significantly the technology to be used.

2.4 The frontier for the French web

Given its mission and since others are doing it for other portions of the web, the BnF wants to focus on the French web. The notion of perimeter is relatively clear for the existing legal deposit (e.g, for books, the BnF requests a copy of each book edited by a French editor). On the web, national borders are blurred and many difficulties arise when trying to give a formal definition of the perimeter. The following criteria may be used:

- The French language. Although this may be determined from the contents of pages, it is not sufficient because of the other French speaking countries or regions e.g. Quebec. Also, many French sites now use English, e.g. there are more pages in English than in French in *inria.fr*.
- The domain name. Resource locators include a domain name that sometimes provides information about the country (e.g. *.fr*). However, this information is not sufficient and cannot in general be trusted. For instance *www.multimania.com* is hosting a large number of French associations and French personal sites and is mostly used by French people. Moreover, the registration process for *.fr* domain names is more difficult and expensive than for others, so many French sites choose other suffixes, e.g. *.com* or *.org*.
- The *address* of the site. This can be determined using information obtainable from the web (e.g., from domain name servers) such as the physical location of the web

server or that of the owner of the web site name. However, some French sites may prefer to be hosted on servers in foreign countries (e.g., for economical reasons) and conversely. Furthermore, some web site owners may prefer to provide an address in exotic countries such as Bahamas to save on local taxes on site names. (With the same provider, e.g., Gandi, the cost of a domain name varies depending on the country of the owner.)

Note that for these criteria, negative information may be as useful as positive ones, e.g., we may want to exclude the domain name *.ca* (for Canada).

The Royal library of Sweden, which has been archiving the Swedish Web for more than 6 years now, has settled on an inclusion policy based on national domain (.se and .nu), checking the physical address of generic domain name owners, and the possibility to manually add other sites. The distribution of the domain names is about 65 percent for nation domains (.se and .nu) and 25 percent for generic domains (.net, .com, .org).

Yet another difficulty in determining the perimeter is that the legal deposit is typically not very interested in commercial sites. But it is not easy to define the notion of commercial site. For instance, *amazon.fr* (note the ".fr") is commercial whereas *groups.yahoo.com/group/vertsdesevres/* (note the ".com") is a public, political forum that may typically interest the legal deposit. As in the case of the language, the nature of web sites (e.g., commercial vs. non commercial) may be better captured using the contents of pages.

No single criteria previously mentioned is sufficient to distinguish the documents that are relevant for the legal deposit from those that are not. This leads to using a multi-criteria based clustering. The clustering is designed to incorporate crucial information: the connectivity of the web. French sites are expected to be tightly connected. Note that here again, this is not a strict law. For instance, a French site on DNA may strongly reference foreign sites such as Mitomap (a popular database on the human mitochondrial genome).

Last but not least, the process should involve the BnF librarians and their knowledge of the web. They may know, for instance, that *00h00.com* is a web book editor that should be archived in the legal deposit.

Technical corner. The following technique is used. A crawl of the web is started. Note that sites specified as relevant by the BnF librarians are crawled first and the relevance of their pages is fixed as maximal. The pages that are discovered are analyzed for the various criteria to compute their *relevance* for the legal deposit. Only the pages believed to be relevant ("suspect" pages) are crawled. For the experiments, the BAO algorithm is used [1] that allows to compute page relevance on-line while crawling the web. The algorithm focuses the crawl to portions of the web that are evaluated as relevant for the legal deposit. This is in spirit of the XML-focused on-line crawling presented in [14], except that we use the multi-criteria previously described. The technique has the other advantage that it is not necessary to store the graph structure of the web and so it can be run with very limited resources. Intuitively, consider L the link matrix of the web (possibly normalized by out-degrees), and X the value vector for any page-based criteria. Then, $L * X$ represents a *depth-1* propagation of the criteria, and in general $L^n * X$ represents the propagation up to depth n. Note that the PageRank [5] is defined by the limit of $L^n{*}X$ when n goes to infinity. We are not exactly interested in PageRank, but only in taking into account some contribution of connectivity. Thus we define the vector value for a page as: $V = X + a_1 * L * X + a_2 * L^2 * X + a_3 * L^3 * X + ...$ Any distribution can be used for the sequence $a_1, a_2, ..., a_n$, as long as the sum converges. When the sequence decreases faster, the contribution of connectivity is reduced.

Since the same technology is used to obtain the *importance* of pages, a more detailed presentation of the technique is delayed to Section 3.

To conclude this section, we note that for the first experiments that we mention in the following sections, the perimeter was simply specified by the country domain name (*.fr*). We intend to refine it in the near future.

3 Building the Archive

In this section, we present a framework for building the archive. Previous work in this area is abundant [15, 2, 11], so we focus on the specificities of our proposal.

A simple strategy would be to take a snapshot of the French web regularly, say n times a year (based on available resources). This would typically mean running regularly a crawling process for a while (a few weeks). We believe that the resulting archive would certainly be considered inadequate by researchers. Consider a researcher interested in the French political campaigns in the beginning of the 21st century. The existing legal deposit would give him access to all issues of the *Le Monde* newspaper, a daily newspaper. On the other hand, the web archive would provide him only with a few snapshots of *Le Monde* web site per year. The researcher needs a more “real time” vision of the web. However, because of the size of the web, it would not be reasonable/feasible to archive each site once a day even if we use sophisticated versioning techniques (see Section 5).

So, we want some sites to be very accurately archived (almost in real-time); we want to archive a very extensive portion of the French web; and we would like to do this under limited resources. This leads to distinguishing between sites: the most important ones (to be defined) are archived frequently whereas others are archived only once in a long while (yearly or possibly never). A similar problematic is encountered when indexing the web [5]. To take full advantage of the bandwidth of the crawlers and of the storage resources, we propose a general framework for building the web archive that is based on a measure of importance for pages and of their change rate. This is achieved by adapting techniques presented in [14, 1]. But first, we define intuitively the notion of importance and discuss the notion of web site.

Page importance. The notion of page importance has been used by search engines with a lot of success. In particular, Google uses an authority definition that has been widely accepted by users. The intuition is that a web page is important if it is referenced by many important web pages. For instance, Le Louvre’s homepage is more important than an unknown person homepage: there are more links pointing to Le Louvre coming from other museums, tourist guides, or art magazines and many more coming from unimportant pages. An important drawback is that this notion is based strictly on the graph structure of the web and ignores important criteria such as language, location and also content.

3.1 Site vs. page archiving

Web crawlers typically work at the granularity of pages. They select one URL to load in the collection of URLs they know of and did not load yet. The most primitive crawlers select the “first” URL, whereas the sophisticated ones select the most “important” URL [5, 14]. For an archive, it is preferable to reason at the granularity of web sites rather than just web pages. Why? If we reason at the page level, some pages in a site (more important than others) will be read more frequently. This results in very poor views of websites. The pages of a particular site would typically be crawled at different times (possibly weeks apart), leading to dangling pointers and inconsistencies. For instance, a page that is loaded may contain a reference to a page that does not exist anymore at the time we attempt to read it or to a page whose content has been updated[5].

For these reasons, it is preferable to crawl sites and not individual pages. But it is not straightforward to define a web site. The notion of web site loosely corresponds to that of editor for the classical legal deposit. The notion of site may be defined, as a first approximation, as the physical site name, e.g., *www.bnf.fr*. But it is not always appropriate to do so. For instance, *www.multimania.com* is the address of a web provider that hosts a large quantity of sites that we may want to archive separately. Conversely, a web site may be spread between several domain names: INRIA’s website is on *www.inria.fr*, *www-rocq.inria.fr*, *osage.inria.fr*, *www.inrialpes.fr*, etc. There is no simple definition. For instance, people will not all agree when asked whether *www.leparisien.fr/news* and *www.leparisien.fr/ shopping* are different sites or parts of the same site. To be complete, we should mention the issue of detecting mirror sites, that is very important in practice.

It should also be observed that site-based crawling contradicts compulsory crawling requirements such as the prevention of *rapid firing*. Crawlers typically balance load over many websites to maximize bandwidth use and avoid over-flooding web servers. In contrast, we focus resources on a smaller amount of websites and try to remain at the limit of rapid firing for these sites until we have a copy of each. An advantage of this focus is that very often a small percentage of pages causes most of the problem. With site-focused crawling, it is much easier to detect server problems such as some dynamic page server is slow or some remote host is down.

3.2 Acquisition: Crawl, Discovery and Refresh

Crawl. The crawling and acquisition are based on a technique [14] that was developed at INRIA in the Xyleme project. The web data we used for our first experiments was obtained by Xyleme [19] using that technology. It allows, using a cluster of standard PCs, to retrieve a large amount of pages with limited resources, e.g. a few million pages per day per PC on average. In the spirit of [7, 8, 14], pages are read based on their importance and refreshed based on their importance and change frequency rate. This results in an optimization problem that is solved with a dynamic algorithm that was

[5] To see an example, one of the authors (an educational experience) used, in the website of a course he was teaching, the URL of an HTML to XML wrapping software. A few months later, this URL was leading to a pornographic web site. (Domain names that are not renewed by owners are often bought for advertisement purposes.) This is yet another motivation for archives.

presented in [14]. The algorithm has to be adapted to the context of the web legal deposit and site-based crawling.

Discovery. We first need to allocate resources between the discovery of new pages and the refreshing of already known ones. For that, we proceed as follows. The size of the French web is estimated roughly. In a first experiment using only ".fr" as criteria and a crawl of close to one billion of URLs, this was estimated to be about 1-2 % of the global web, so of the order of 20 millions URLs. Then the librarians decide the portion of the French web they intend to store, possibly all of it (with all precautions for the term "all"). It is necessary to be able to manage in parallel the discovery of new pages and the refresh of already read pages. After a stabilization period, the system is aware of the number of pages to read for the first time (known URLs that were never loaded) and of those to refresh.

It is clearly of interest to the librarians to have a precise measure of the size of the French web. At a given time, we have read a number of pages and some of them are considered to be part of the French web. We know of a much greater number of URLs, of which some of them are considered "suspects" for being part of the French web (because of the ".fr" suffix or because they are closely connected to pages known to be in the French web, or for other reasons.) This allows us to obtain a reasonably precise estimate of the size of the French web.

Refresh. Now, let us consider the selection of the next pages to refresh. The technique used in [14] is based on a cost function for each page, the penalty for the page to be stale. For each page p, $cost(p)$ is proportional to the importance of page $i(p)$ and depends on its estimated change frequency $ch(p)$. We define in the next subsection the importance $i(S)$ of a site S and we also need to define the "change rate" of a site. When a page p in site S has changed, the site has changed. The change rate is, for instance, the number of times a page changes per year. Thus, the upper bound for the change rate of S is $ch(S) = \sum_{p\ in\ S}(ch(p))$. For efficiency reasons, it is better to consider the average change rate of pages, in particular depending on the importance of pages. We propose to use a weighted average change rate of a site as:

$$c\bar{h}(S) = \frac{\sum_p ch(p) * i(p)}{\sum_p i(p)}$$

Our refreshing of web site is based on a cost function. More precisely, we choose to read next the site S with the maximum ratio:

$$\rho(S) = \frac{\theta(i(S), c\bar{h}(S), lastCrawl(S), currentTime)}{\text{number of pages in } S}$$

where θ may be, for instance, the following simple cost function:

$$\theta = i(S) * (currentTime - lastCrawl(S)) * c\bar{h}(S)$$

We divide by the number of pages to take into account the cost to read the site. A difficulty for the first loading of a site is that we do not know for new sites their number of pages. This has to be estimated based on the number of URLs we know of the site (and never read). Note that this technique forces us to compute importance at page level.

To conclude this section, we will propose a model to avoid such an expensive computation. But first we revisit the notion of importance.

3.3 Importance of pages for the legal deposit

When discovering and refreshing web pages, we want to focus on those which are of interest for the legal deposit. The classical notion of importance is used. But it is biased to take into account the perimeter of the French web. Finally, the content of pages is also considered. A librarian typically would look at some documents and know whether they are interesting. We would like to perform such an evaluation automatically, to some extent. More precisely, we can use for instance the following simple criteria:

- **Frequent use of infrequent Words:** The frequency of words found in the web page is compared to the average frequency of such words in the French web[6]. For instance, for a word w and a page p, it is:

  $$I_w = \sum_{each\ word} \frac{f_{p,w}}{f_{web}} \quad \text{where } f_{p,w} = n_{p,w}/N_p$$

  and $n_{p,w}$ is the number of occurrences of a word w in a page and N_p the number of words in the page. Intuitively, it aims at finding pages dedicated to a specific topic, e.g. butterflies, so pages that have some content.
- **Text Weight:** This measure represents the proportion of text content over other content like HTML tags, product or family names, numbers or experimental data. For instance, one may use the number of bytes of French text divided by the total number of bytes of the document.

  $$I_{pt} = \frac{size_{french\ words}}{size_{doc}}$$

  Intuitively, it increases the importance of pages with text written by people versus data, image or other content.

A first difficulty is to evaluate the relevance of these criteria. Experiments are being performed with librarians to understand which criteria best match their expertise in evaluating sites. Another difficulty is to combine the criteria. For instance, *www.microsoft.fr* may have a high PageRank, may use frequently some infrequent words and may contain a fair proportion of text. Still, due to its commercial status, it is of little interest for the legal deposit. Note that librarians are vital in order to “correct” errors by positive action (e.g., forcing a frequent crawl of *00h00.com*) or negative one (e.g., blocking the crawl of *www.microsoft.fr*). Furthermore, librarians are also vital to correct the somewhat brutal nature of the construction of the archive. Note however that because of the size of the

[6] To guarantee that the most infrequent words are not just spelling mistake, the set of words is reduced to words from a French dictionary. Also, as standard, stemming is used to identify words such as *toy* and *toys*.

web, we should avoid as much as possible manual work and would like archiving to be as fully automatic as possible.

As was shown in this section, the quality of the web archive will depend on complex issues such as being able to distinguish the borders of a web site, analyze and evaluate its content. There are ongoing projects like THESU [6] which aim at analyzing thematic subsets of the web using classification, clustering techniques and the semantics of links between web pages. Further work on the topic is necessary to improve site discovery and classification

To conclude this section, we need to extend previously defined notions to the context of website. For some, it suffices to consider the site as a huge web document and aggregate the values of the pages. For instance, for *Frequent use of infrequent Words*, one can use:

$$I_w = \sum_{each\ word} \frac{f_{site}}{f_{web}} \quad \text{where} \;\; f_{S,w} = \sum_{p\ in\ S} (n_{p,w}) / \sum_{p\ in\ S} (N_p)$$

Indeed, the values on word frequency and text weight seem to be more meaningful at the site level than at the page level.

For page importance, it is difficult. This is the topic of next section.

4 Site-based Importance

To obtain a notion of site importance from the notion of page importance, one could consider a number of alternatives:

- Consider only links between websites and ignore internal links;
- Define site importance as the sum of PageRank values for each page of the web site;
- Define site importance as the maximum value of PageRank, often corresponding to that of the site main page.

We propose in this section an analysis of site importance that will allow us to choose one notion.

First, observe that the notion of page importance is becoming less reliable as the number of dynamic pages increases on the web. A reason is that the semantics of the web graph created by dynamic pages is weaker than the previous document based approach. Indeed, dynamic pages are often the result of database queries and link to other queries on the same database. The number of incoming/outgoing links is now related to the size of the database and the number of queries, whereas it was previously a human artifact carrying stronger semantics. In this section, we present a novel definition of sites' importance that is closely related to the already known page importance. The goal is to define a site importance with stronger semantics, in that it does not depend on the site internal databases and links. We will see how we can derive such importance from this site model.

Page importance, namely PageRank in Google terminology, is defined as the fixpoint of the matrix equation $X = L * X$ [18, 16], where the web-pages graph G is represented as a link matrix $L[1..n, 1..n]$. Let $out[1..n]$ be the vector of out-degrees.

If there is an edge for i to j, $L[i,j] = 1/out[i]$, otherwise it is 0. We note $I_{page}[1..n]$ the importance for each page. Let us define a web-sites graph G' where each node is a web-site (e.g. *www.inria.fr*). The number of web-sites is n'. For each link from page p in web-site Y to page q in web-site Z there is an edge from Y to Z. This edges are weighted, that is if page p in site S is twice more important than page p' (in S also), then the total weight of outgoing edges from p will be twice the total weight of outgoing edges from p'. The obvious reason is that browsing the web remains page based, thus links coming from more important pages deserve to have more weight than links coming from less important ones. The intuition underlying these measures is that a web observer will visit randomly each page proportionally to its importance. Thus, the link matrix is now defined by:

$$L'[Y,Z] = \sum_{p\ in\ Y,\ q\ in\ Z} \frac{I_{page}[p]}{\sum_{p'\ in\ Y} I_{page}[p']} * L[p,q]$$

We note two things:

– If the graph G representing the web-graph is (artificially or not) strongly connected, then the graph G' derived from G is also strongly connected.
– L' is still a stochastic matrix, in that $\forall Y, \sum_Z L'[Y,Z] = 1$. (proof in appendix).

Thus, the page importance, namely PageRank, can be computed over G', L' and there is a unique fixpoint solution. We prove in appendix that the solution is given by:

$$I_{site}[Y] = \sum_{p\ in\ Y} I_{page}[p]$$

This formal relation between website based importance and page importance suggests to compute page importance for all pages, a rather costly task. However, it serves as a reference to define site-based importance, and helps understand its relation to page-based importance. One could simplify the problem by considering, for instance, that all pages in a website have the same importance. Based on this, the computation of site-importance becomes much simpler. In this case, if there is there is at least one page in Y pointing to one page in Z, we have $L'[Y,Z] = 1/out(Y)$, where $out(Y)$ is the out-degree of Y. A more precise approximation of the reference value consists in evaluating the importance of pages of a given website S on the restriction of G to S. Intuitively it means that only internal links in S will be considered. This approximation is very effective because: (i) it finds very good importance values for pages, that correspond precisely to the internal structure of the web-site (ii) it is cheaper to compute the internal page importance for all websites, one by one, than to compute the PageRank over the entire web (iii) the semantics of the result are stronger because it is based on site-to-site links.

This web-site approach enhances significantly previous work in the area, and we will see in next section how we also extend previous work in change detection, representation and querying to web sites.

5 Representing Changes

Intuitively, change control and version management are used to save storage and bandwidth resources by updating in a large data warehouse only the small parts that have changed [10]. We want to maximize the use of bandwidth, for instance, by avoiding the loading of sites that did not change (much) since the last time they were read. To maximize the use of storage, we typically use compression techniques and a clever representation of changes. We propose in this section a change representation at the level of web sites in the spirit of [9, 10]. Our change representation consists of a **site-delta**, in XML, with the following features:

(i) Persistent identification of web pages using their URL, and unique identification of each document using the tuple (URL, date-of-crawl);

(ii) Information about mirror sites and their up-to-date status;

(iii) Support for temporal queries and browsing the archive

The following example is a **site-delta** for *www.inria.fr*:

```
<website url="www.inria.fr">
<page url="/index.html">
  <document date="2002-Jan-01" status="updated"
          file="543B6.html"/>
  <document date="2002-Mar-01" status="unchanged"
          file="543B6.html"/>
</page>
<page url="/news.html">
  <document date="2002-Mar-25" status="updated"
          file="543GX6.html"/>
  <document date="2002-Mar-24" status="error">
    <error httperror="404"/>
  </document>
  <document date="2002-Mar-23" status="updated"
          file="523GY6.html"/>
  ...
  <document date="1999-Jan-08" status="new"
          file="123GB8.html"/>
</page>
<mirror url="www-mirror.inria.fr" depth="nolimit">
  <exclusion path="/cgi-bin" />
</mirror>
</website>
```

Each web-site element contains a set of pages, and each page element contains a subtree for each time the page was accessed. If the page was successfully retrieved, a reference to the archive of the document is stored, as well as some metadata. If an error was encountered, the page status is updated accordingly. If the page mirrors another page on the same (or on another) web-site, the document is stored only once (if possible) and is tagged as a mirror document. Each web-site tree also contains a list of web-sites mirroring part of its content. The up-to-date status of mirror sites is stored in their respective XML file.

Other usages. The site-delta is not only used for storage. It also improves the efficiency of the legal deposit. In particular, we mentioned previously that the legal deposit works at a site level. Because our site-delta representation is designed to maintain information at page level, it serves as an intermediate layer between site-level components and page-based modules.

For instance, we explained that the acquisition module crawls sites instead of pages. The site-delta is then used to provide information about pages (last update, change frequency, file size) that will be used to reduce the number of pages to crawl by using caching strategies. Consider a news web site, e.g. *www.leparisien.fr/*. News articles are added each day and seldom modified afterwards, only the index page is updated frequently. Thus, it is not desirable to crawl the entire web site every day. The site-delta keeps track of the metadata for each pages and allows to decide which pages should be crawled. So it allows the legal deposit to virtually crawl the entire web site each day.

Browsing the archive. A standard first step consists in replacing links to the Internet (e.g. *http://www.yahoo.fr/*) by local links (e.g. to files). The process is in general easy, some difficulties are caused by pages using java-scripts (sometimes on purpose) that make links unreadable. A usual problem is the consistency of the links and the data. First, the web graph is not consistent to start; broken links, servers down, pages with out of date data are common. Furthermore, since pages are crawled very irregularly, we never have a true snapshot of the web.

The specific problem of the legal deposit is related to *temporal browsing*. Consider, for instance, a news web site that is entirely crawled every day. A user may arrive at a page, perhaps via a search engine on the archive. One would expect to provide him the means to browse through the web site of that day and also in time, move to this same page the next day. The problem becomes seriously more complex when we consider that all pages are not read at the same time. For instance, suppose a user reads a version t of page p and clicks on a link to p'. We may not have the value of page p' at that time. Should we find the latest version of p' before t, the first version after t, or the closest one? Based on an evaluation of the change frequency of p', one may compute which is the most likely to be the correct one. However, the user may be unsatisfied by this and it may be more appropriate to propose several versions of that page.

One may also want to integrate information coming from different versions of a page into a single one. For instance, consider the index of a news web site with headlines for each news article over the last few days. We would like to *automatically* group all headlines of the week into a single index page, as in Google news search engine [4]. A difficulty is to understand the structure of the document and to select the valuable links. For instance, we don't want to group all advertisements of the week!

6 Conclusion

As mentioned in the introduction, the paper describes preliminary work. Some experiments have already been conducted. A crawl of the web was performed and data is now being analyzed by BnF librarians. In particular, we analyze the relevance of page importance (i.e., PageRank in Google terminology). This notion has been to a certain extent validated by the success of search engines that use it. It was not clear whether it is adapted to web archiving. First results seem to indicate that the correlation between our automatic ranking and that of librarians is essentially as similar as the correlation between ranking by librarians.

Perhaps the most interesting aspect of this archiving work is that it leads us to reconsider notions such as web site or web importance. We believe that this is leading us to a better understanding of the web. We intend to pursue this line of study and try to see how to take advantage of techniques in classification or clustering. Conversely, we intend to use some of the technology developed here to guide the classification and clustering of web pages.

Acknowledgments We would like to thank Laurent Mignet, Benjamin Nguyen, David Leniniven and Mihai Preda for discussions on the topic.

References

[1] S. Abiteboul, M. Preda, and G. Cobena. Computing web page importance without storing the graph of the web (extended abstract). In *IEEE Data Engineering Bulletin, Volume 25*, 2002.

[2] A. Arvidson, K. Persson, and J. Mannerheim. The kulturarw3 project - the royal swedish web archiw3e - an example of ’complete’ collection of web pages. In *66th IFLA Council and General Conference*, 2000. www.ifla.org/IV/ifla66/papers/154-157e.htm.

[3] M.K. Bergman. The deep web: Surfacing hidden value. www.brightplanet.com/.

[4] Google. Google news search. http://news.google.com/.

[5] Google. www.google.com/.

[6] Maria Halkidi, Benjamin Nguyen, Iraklis Varlamis, and Mihalis Vazirgianis. Thesus: Organising web document collections based on semantics and clustering. Technical Report, 2002.

[7] T. Haveliwala. Efficient computation of pagerank. *Technical report, Stanford University*, 1999.

[8] H. Garcia-Molina J. Cho. Synchronizing a database to improve freshness. *SIGMOD*, 2000.

[9] R. Lafontaine. A delta format for XML: Identifying changes in XML and representing the changes in XML. In *XML Europe*, 2001.

[10] A. Marian, S. Abiteboul, G. Cobena, and L. Mignet. Change-centric management of versions in an XML warehouse. *VLDB*, 2001.

[11] L. Martin. Networked electronic publications policy, 1999 . www.nlc-bnc.ca/9/2/p2-9905-07-f.html.

[12] J. Masanes. Préserver les contenus du web. In *IVe journées internationales d’études de l’ARSAG - La conservation à l’ère du numérique*, 2002.

[13] J. Masanès. The BnF’s project for web archiving. In *What’s next for Digital Deposit Libraries? ECDL Workshop*, 2001 . www.bnf.fr/pages/infopro/ecdl/france/sld001.htm.

[14] L. Mignet, M. Preda, S. Abiteboul, S. Ailleret, B. Amann, and A. Marian. Acquiring XML pages for a WebHouse. In *proceedings of Base de Données Avancées conference*, 2000.

[15] A National Library of Australia Position Paper. National strategy for provision of access to australian electronic publications. www.nla.gov.au/policy/paep.html.

[16] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1998.

[17] S. Raghavan and H. Garcia-Molina. Crawling the hidden web. In *The VLDB Journal*, 2001.

[18] L. Page S. Brin. The anatomy of a large-scale hypertextual web search engine. *WWW7 Conference, Computer Networks 30(1-7)*, 1998.

[19] Xyleme. `www.xyleme.com`.