World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme: http://www.ifla.org/IV/ifla71/Programme.htm

June 9, 2005

Code Number: 074-E
Meeting: 133 SI - Bibliography

Web crawling : The Bibliothèque nationale de France experience

Christian Lupovici
Head of the National Bibliographic Agency
Bibliothèque nationale de France
Christian.lupovici@bnf.fr

Abstract:
The Bibliothèque nationale de France in the framework of its Legal deposit mission, is currently experimenting with web harvesting procedures and is organising the long-term preservation of digital documents. The work carried out to achieve this goal includes fundamental thoughts on the essence of Legal deposit and on the bibliographic treatment of Internet resources. Working on real scale, the huge amount of digital resources is an important factor to help in any decision to be taken by the National Library.

1. Legal deposit at the BnF

1.1 Historic background

The French Legal deposit legislation goes back to the sixteenth century when François the 1st took the Ordonnance de Montpellier in 1537. Since that date, the purpose of all posterior regulations was to take into account each new technology arising and to integrate it into the legal provisions making the legal deposit scope broader each time.

1537: books;
1667: etchings;
1689: engravings;
1925: any graphic art production;
1941: posters, scores and photography;
1963: sound recordings of any nature;
1975: records of still or moving pictures, whatever the technical process is;
1977: Cinema works;
1992: All kinds of documents on media, including databases and expert systems as far as they are distributed to the public.

The inclusion of web resources into the French regulation will happen in 2005, as the matter is being discussed in Parliament.

1.2 Mission and scope of Legal deposit

It is clear that the purpose of Legal deposit is to preserve and give access for the long term to the whole cultural heritage of documents produced and distributed in numerous copies and to give access to that heritage. Legal deposit includes all sorts of resources and is not restricted only to the French language, to only the documents published on the national territory.

It is more a question of type of content and/or distribution channels (books, engravings, maps, sounds, audiovisuals) than a question of media: paper, disc or online. On the Internet, most of institutional grey literature will be part of the Legal deposit regulation as long as documents on the web are accessible to the public, thus published.

Legal deposit means also that the collecting is not a question of acquisition, meaning a collection development policy, but a question of collecting material of all sorts that gives a picture of contents distributed to French people in the present time, in order to preserve it for future historical studies. This means that all different versions of a document as well as forbidden works, are included, as they have always been since the sixteenth century.
How to transpose this to the Web resources?

2. The Web crawling

2.1 The Internet harvesting for Legal deposit purpose

The transposition of this Legal deposit philosophy to the Web resources implies:

a) to put out of scope all domain names that are not obviously containing resources belonging to French culture;
b) to focus on the specific French national domain names (*.fr, .pf, .wf, .pm, .re, .tf, .ad, .yt*);
c) to harvest all the rest.

In a first approach, it is not possible to know what is relevant for French culture within "the rest" and a work has to be carried out to find a process which make a rough selection based on a semantic analysis on the sites' content. Nevertheless, as the storage space is not anymore a crucial problem, the Bibliothèque nationale de France is storing and preserving a set of sites much larger than what would be strictly necessary for the Legal deposit purpose, waiting for better solution. In any case, the part of the web that will be stored and preserved will always be large.

2.2 From experiments to real scale

2.2.1 First experiments

The Bibliothèque nationale de France has been collecting French web sites managing HTTrack crawler since 2001[1]. The first experiments were focused on several campaigns for elections in France. These experiments were aims at testing different points:
a) how to build a typology of web sites. And according to it what is the right periodicity for crawling;
b) to test technical problems (formats, depth of the crawl) and tune the crawler;
c) to assess manual and automatic mechanisms of web sites selection.
This phase was important to assess automatic selection of web sets to be processed manually and how to use a tool to help in decision making.

2.2.2 Web crawling on real scale

What is the French Legal deposit in terms of resources and volume collected?
An extract of a snapshot made on the whole web during the spring 2004 by Alexa web crawler was produced, according to the specifications on the domains to be crawled (*.fr, .com, .net, .edu, .com, .biz, .info, .int, .pf, .ad, .coop, .name, .aero, .tf, .re, .museum, .pro, .pm, .wf, .yt*) and

[1] http://bibnum.bnf.fr/conservation/aristote2004/roche.pdf

those to be excluded. The result of this crawl gave back more than 2 billion URLs and 65 million hosts, weighting 27 terabytes.

In addition, a focused crawl of the *.fr* domain was carried out at the end of 2004 and January 2005 using the Heritrix crawler. The result of this latter crawl gave back 500 000 public sites and 4 000 personal sites. That is more than 118 million URLs and a half million hosts, weighting 3 terabytes.

For this volume of data, it is useless to imagine any manual cataloguing or indexing process other than at the fringe. Such amount of data would take a minimum of 500 man/years to be sorted and catalogued only for the plubic sites. This is twice more than the cataloguing staff for the National bibliography at BnF.

This is why it is necessary to combine different techniques.

2.3 Web Legal deposit processing

The fundamental method is an automatic crawling. The entire surface web and some of the deep web can be harvested in such a way, depending on the technical problem the crawler is faced with and how it has been tuned. But generally speaking, an automatic crawl will mainly collect the surface web, creating a map of the web.

A second and complementary method is to focus the automatic crawling on a particular set of sites, for instance, on the *.fr* domain, and ask the crawler for a focussed harvesting in depth. These two last methods are top down oriented.

In order to get the whole web picture, we must combine the periodic automatic harvesting, with a bottom up oriented technique which consists in complementing the collections with deep web resources deposited by sites producers and to recreate the links with the surface web already harvested. Those deposits will have to be processed partly manually, partly automatically. The deposit can present advantages in term of preservation insurance, but it's a time consuming process, starting by the analysis of the site, which takes 5 man/days on average.

The whole set of data is stored at the BnF in a central data repository for internal use and staff appropriation. The web archive is part of the global BnF digital storage and preservation programme, SPAR (*Système de Préservation des ARchives*) launched at the end of 2004.

2.4 Question and method of selection

In principle, the Bibliothèque nationale de France would like to archive the most relevant web sites. So, it is interesting to find a method that can select or propose a selection among the whole sites harvested. A French research project named "Watson" was carried out on linguistic treatments to be applied on the BnF web archive for users and staff. The project lasted from 2003 to 2004[2]. One of the ways it explored was the characterisation of the site using linguistic analysis of the metadata and the phrases included in the full text. This method provided abstracts on the site content. It is not yet good enough to be used without any human checking, but it is good enough to be a helping tool for site selection.

The BnF is organising a web watching team composed of professional staff working in the different Library departments in charge of legal deposit, mainly. The team's objective is to learn how to deal with this new digital resource environment and make the necessary connection with previous traditional collections on media. Access tools and watch methods have been set up to assess the result of the crawl and the quality of the harvest. The team has access to a tool that enables them to make verifications both in the BnF archive and on the web, in order to propose specific web sites to be crawled in depth through a focused crawl. This fall, the crawl will take into account the work achieved by the watching team and will have the proposed URLs "seeds" included into its parameters.

[2] José Coch, Julien Masanès**. -** Language engineering techniques for web archiving. In : 4th International Web Archiving Workshop (IWAW04), 16 septembre 2004, Bath, UK - http://www.iwaw.net/04/index.html (visited 30/05/2005).

Anyway, it will never be possible to select a set of web sites corresponding exactly to the French Legal deposit and it is obvious that the BnF web archive will always be of a larger size. If all countries are doing the same, there will be overlapping, but it is rather a good thing as storage is not anymore a problem and duplicates in different countries will ensure a better security for long term preservation. On another hand, the snapshot technique does not ensure any comprehensiveness of the collections but a representative sample of the actual culture. That is the essence of Legal deposit.

Nevertheless, a need for a "smart crawler" able to focus on site sets according to given parameters and able to go in depth when the rank of relevance is high, has been recognised by the international consortium IIPC (International Internet Preservation Consortium). The British Library and the Bibliothèque nationale de France have launched, this spring, a call for proposal for a smart crawler.

2.5 New environment, new document

Dealing with the web we must leave the traditional document approach for processing and move toward a new way of managing the resources.

The web archive has new dimensions:
It must keep its navigability that is recreated internally. None of the documents is a stand-alone resource. The document itself is being transformed. In the BnF model, the documentary unit is the web site itself (in a collection model).

2.6 Linguistic techniques for resource analysis

Facing the huge amount of data ingested in the storage system, both staff and readers need tools in order to search information they need. They may need to gather resources in subsets looking after precise information, but they also may need to process a big amount of information in order to apply statistic or linguistic treatments.

The Watson project aimed at selecting customised sets of resources corresponding to a designated domain, and to apply on this identified set some linguistic analysis.
It may give an idea of the content by telling the user about the structure, keywords, names, and locations etc. embedded in the documents or in giving a navigable abstract.
The project produced user specific tools for subject extraction, navigation in context and semantic analysis of the content of sites. The experiments were carried out on the French elections archive.

3. New cataloguing prospective

3.1 Cataloguing metadata

Traditionally metadata have been used for describing, identifying physical objects for retrieval purposes. This cataloguing activity is labour intensive work and totally manual, but it is adapted to physical objects and can't be avoided in this context.

In a digital environment, the situation is different:

Digital objects are (or should be) ready for information processing. They contain self-documentation, which can be technical and legal metadata enabling their management in a system.

When text is involved, the content itself can be indexed, sorted, without any human intervention (at this stage). This covers structure and semantic content information.
In the web context, if descriptive metadata can be poor, they are balanced by the network effect. The digital objects are non-isolated documents in the web archive. They are linked to other resources by explicit links, content and description of content, which include finding aids--whether traditional or totally new.

This does not mean that we do not need anymore descriptive metadata or keyword. But the library in not the best place to create this information, it is better to transfer this task to the

creation phase and the need for this type of data is less important than it is in a traditional environment.

3.2 Other kinds of metadata

New issues are emerging, like rights management, like ensuring authenticity, not only for legal reasons, but also to ensure infinite reproduction of digital resources and their instantaneous circulation. Formats, which are often combined in a single document, are one of the crucial issues, as the server's or the document's information is not reliable enough. Documents collected need a format assessment.

In the context of a National Library, it is also very important to extend metadata to the preservation information adapted to a web archive.

3.3 Cataloguing future

The cataloguing situation in the web context can be summarised as making specifications for digital document creation and resource analysis during the ingest phase (OAIS model), including the crawler instructions and the interaction with web producers.

4. What a National Bibliography of the web can be?

The French Legal deposit law (even the new one) says that there will be a National Bibliography made from the document deposited.

But in this context, what does a National Bibliography look like?

As the purpose of a National Bibliography is to advertise new publications available, we can envisage to publish on the BnF web site, the list of new sites coming in (not the modifications), applying an automatic sorting and "cataloguing" treatment at the archiving phase (OAIS model). The result will be a list of URLs with name index, subject index (as far as it is possible to make it with a linguistic analysis, close to what has been done in the Watson project).