International Workshop on digital preservation and copyright
At WIPO, Geneva
July 15, 2008

Topic 5
Preservation of Web-based materials

The French perspective

Christian Lupovici
Bibliothèque nationale de France

France benefit from a very old and well established legislation on Legal Deposit which goes back to the sixteenth century, starting with books in 1537 and adding supplementary regulations to take into account each time a new technology appeared[1], up to 1992 when the last regulation stated "all kinds of documents on media" as far as they are distributed to the public.

For information made available online (specially via the internet), there was no addition to the Legal deposit law but the permission to archive and preserve the content of the internet as an exception to the law on Authors' rights and neighbouring rights in the Information society published in August 2006. In fact, this law encompasses 3 exceptions to the authors' rights:

a) an exception for libraries allowing them to make a copy of a digital document for preservation;
b) an exception for handicapped people to make possible adaptations from the publishers digital files;
c) an exception for Legal deposit to make possible the web archiving.

The National library of France has in fact anticipated the publication of the law, harvesting and archiving the web in collaboration with Internet Archive since the year 2000.
As the web is harvested in the framework of Legal deposit, there is no permission to ask – this is the great advantage of having a Legal deposit law – but the crawler identifies itself as the Legal deposit crawler.
Up to the publication of the law, the crawler respected the "robot.txt" exclusion. It is not anymore the case as it became fully legal.

For the materials which are not captured by crawling, for technical reasons, the regulation states that the Legal deposit institution is allowed to ask the publisher to come to a deposit protocol. Of course all materials submitted to Legal deposit must be TPM free for preservation.

The scope of the harvesting tends to be similar to the physical Legal deposit. That means as comprehensive as possible and including all kind of type of document (text, image, maps, sound, video, blogs…) as the information is accessible by the public. Nevertheless this demand to put in place technical parameters to target the French Legal deposit scope, specially when the crawler leaves the French top level domains and is going to generic top level domains following the links.

[1] Christian Lupovici. - Web crawling: the bibliothèque nationale de France experience. 71st IFLA general conference and council, August 2005, Oslo, Norway. http://www.ifla.org/IV/ifla71/papers/074e-Lupovici.pdf

BnF crawls twice a year on a large scale (Legal deposit scope) and manages several supplementary focused crawls depending on special events or topics like elections campaigns.

The whole harvested materials is archived without any selection or cataloguing process. The Bibliothèque nationale de France is currently building a trusted repository based on the OAIS model and with back up on a remote site. The system called SPAR[2] encompasses a module which is managing the rights and authorization of access to the materials whether these are coming from the Web or deposits or digitization programmes. The system is using bibliographic data from the catalogue and authority files data in order to have an automatic process when somebody is requesting a document.
If some materials will be lost in the long term for technical reasons, no information is ever deleted by choice from the repository as it has always been.

Readers can officially access the content of the web archive since April 2008 on a trial basis. The older part of the archive can be retrieved by URL through the Wayback Machine and the more recent part of the archive is fully indexed and can be searched by keywords.

The access to the web information is subject to restrictions by regulation, on the same model as the current Legal deposit material on hard copy. It is stated that the materials is only accessible for research to readers registered by the Library. The access is also restricted to the Library buildings.
What seems quite simple is complicated by the big difference in nature of the web-based materials. Therefore, there is an additional restriction to the use of the web archives which relates to the protection of privacy. Although no information is hidden, nor prohibited to access, no use of personal routine to process for data mining, this huge amount of information is allowed. There is an obligation to use only the Library's routines.

Accessing the web-based materials in the Bibliothèque nationale de France is a too short experiment to have a report on the copyright problems we may face in this new environment, but as the library staff have experience in such legal issues we are very confident.

[2] E. Bermès, T. Ledoux, I. Dussert-Carbone, C. Lupovici. – Digital preservation at the National Library of France: a technical and organizational overview. 74th IFLA general conference and council, August 2008, Québec, Canada. http://www.ifla.org/IV/ifla74/papers/084-Bermes_Carbone_Ledoux_Lupovici-en.pdf