The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future

Peter Stirling and **Gildas Illien**
(Main authors)

Pascal Sanz and **Sophie Sepetjan**
(Contributing authors, speakers at IFLA Conference)
Bibliothèque nationale de France
Paris, France

Meeting: **193 — *e-Legal deposit: from legislation to implementation; from ingest to access* — Bibliography Section with IFLA-CDNL Alliance for Digital Strategies Programme (ICADS), Information Technology, National Libraries and Knowledge Management**

Abstract:

The article describes the legal situation in France regarding the legal deposit of digital material, and shows how it has been implemented in practice at the Bibliothèque nationale de France (BnF). The focus is on web archiving, where the BnF has experience going back almost ten years, but other aspects of digital legal deposit are discussed, with possible future developments and challenges. Throughout comparisons are made with the situations in other countries.

The legal deposit of online electronic publications is a relatively recent development, but it is one which takes its place in a long-established tradition of legal deposit legislation in France. This article will demonstrate that digital legal deposit is a natural continuation and evolution of the existing legal situation, while at the same time creating new challenges and demanding the re-examination of some received ideas regarding legal deposit. It seeks to present the legal situation in France and the way in which it is put into practice; while the responsibility for legal deposit is divided between several institutions, this article will concentrate particularly on the Bibliothèque nationale de France (BnF).

The article starts with a brief summary of the history of legal deposit legislation in France, which establishes the aims and spirit of legal deposit legislation. The specific laws and regulations governing legal deposit, and notably legal deposit of electronic publications, are then outlined. The main part of the article then discusses the specific aspects of law and practice in four areas: the acquisition, conservation and description of documents and the means of access to them. For each section the legal possibilities and restrictions are put in the context of actual practice; comparisons are made with the situation in other countries, and

there is a discussion of open questions and future challenges. The conclusion sums up the current situation and suggests ways in which it may develop.

I. The history and background of digital legal deposit in France

Legal deposit in France was created in 1537 by King Francis1st, in what is known as the "Ordonnance de Montpellier". This text obliged printers and booksellers to deposit a copy of every printed book published or made available in France to the Royal Library, which was later to become the National Library. Over the centuries, several legal texts have been put in place to regulate legal deposit, and the legislation has evolved to cover different publication types and forms, hence adjusting to all major technological and social changes. This is particularly true during the 20th and 21st century, when the development of many media innovations created many new forms of publication, which have gradually been included in the scope of legal deposit legislation. The most recent addition, following the 2006 law on Authors' Rights and Related Rights in the Information Society, is electronic publications and the Internet.

History of legal deposit in France

Printed material	1537
Prints, maps and plans	1648
Sheet music	1793
Photographs and sound recordings	1925
Posters	1941
Videos and multimedia documents	1975
Cinema	1977
Multimedia, software and databases	1992
Internet	2006

The idea that the aim of legal deposit is to safeguard the cultural heritage of the country is present from the beginnings; the wording of the 1537 "Ordonnance de Montpellier" shows that the idea of safeguarding books from being lost to posterity is already central. It is true that other aims have been suggested, more or less officially, for legal deposit, such as state control over publications, and protection of copyright. In the former case, legal deposit is sometimes considered as being primarily a matter of state control over what is published: this is not entirely accurate, particularly since in the early years of legal deposit there were already censorship laws in place which assured the state control of publications more effectively than legal deposit. Over time however, the perceived purpose of legal deposit has shifted, with aspects of state control mixed with those of cultural heritage, while the status of a work held under legal deposit has also been used to safeguard copyright, during the period 1793-1925. Since 1925 legal deposit in France no longer plays this role, and today the Code de la propriété intellectuelle (the French Copyright Act), following the Convention of Berne,

specifies that copyright is inherent in published works[1]. However in other countries, notably the United States, legal deposit remains tightly linked with copyright legislation[2].

The heritage basis of legal deposit was affirmed in a revision of the relevant law in 1992, in which the clauses relating to legal deposit were added to the "Code du Patrimoine", the collection of French legislation relating to cultural heritage[3]. The cultural role of legal deposit is also taken up in the decree defining the foundation of the new Bibliothèque nationale de France, dating from 1994. Here the first two missions of the library are defined as:

1) To collect, catalogue, conserve and enrich, in all areas of knowledge, the national heritage for which it has responsibility, in particular the heritage of the French language and French civilisation;
2) To ensure access by the greatest possible number to the collections, with the exception of secrets protected by law, under conditions respecting the legislation on intellectual property and compatible with the conservation of the collections.[4]

It is further made clear in the text that legal deposit is one of the means by which these missions may be fulfilled. As founding missions of the library, these clauses illustrate the spirit of legal deposit, which applies equally to electronic publications as to all other material: legal deposit must collect all material published in France regardless of content, language or value, must preserve it without limit of time, and must make it available to the public, but in conditions which respect intellectual property and which do not pose a risk to the conservation of the material.

The relevant articles of the Code du Patrimoine, along with several other texts, control the manner in which material is collected, conserved and made available. The precise way in which this legislative framework may be applied to electronic publications is detailed in the next section.

II. The legislation governing digital legal deposit in France today

a) Code du Patrimoine, incorporating Legal Deposit Law (1992) and DADVSI (2006)

The principal text governing legal deposit in France is the Code du Patrimoine; in the discussion of various aspects of legal deposit in the course of this article, reference will be made regularly to the different articles of Title III, dedicated to legal deposit. In French law, a Code is a compilation of different laws and regulations in a specific area, and the articles on

[1] *Code de la propriété intellectuelle*, article L111-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006278868&cidTexte=LEGITEXT000006069414&dateTexte=20110520 (consulted 20 May 2011)
[2] *Copyright Law of the United States of America and Related Laws Contained in Title 17 of the United States Code*; Chapter 4: Copyright Notice, Deposit, and Registration; Article 407. Deposit of copies or phonorecords for Library of Congress. http://www.copyright.gov/title17/92chap4.html#407 (consulted 20 May 2011)
[3] *Code du patrimoine*, article L131-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845515&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)
[4] *Décret n°94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France*, article 2. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082797&dateTexte=20110520 (consulted 20 May 2011)

legal deposit and their integration into the Code du Patrimoine come mainly from the Legal Deposit Law, passed in 1992. Digital legal deposit, however, was created with the 2006 law on Authors' Rights and Related Rights in the Information Society (in French, "Droits d'auteur et droits voisins dans la société de l'information", known as DADVSI)[5]. This law is a transposition of the 2011 European Union Copyright Directive (2001/29/CE)[6]. It introduces the possibility of electronic legal deposit as an exception to Copyright for the National Library. Because of its origins (a European directive), this Act has similarities with other pieces of legislation to be found in other European countries, such as Finland or Denmark for instance. As a result, the legal situation described here is not unique to France and may be regarded as fairly representative of other national legislations applicable in Europe, although differences are to be observed from one country to the other.

In the article defining the list of publication subject to legal deposit, the DADVSI introduced the following sentence:

> Also subject to legal deposit are signs, signals, writings, images, sounds or messages of any kind communicated to the public by electronic means.[7]

The definition of electronic publications is phrased in deliberately general terms, to avoid limiting the legislation to specific technologies which may soon become obsolete. The legislation therefore permits, and indeed requires, the legal deposit of everything published on the Internet, while excluding private correspondence (emails, intranets, the private areas of social networks…). This may range from websites in a general sense, to video and sound recordings, or any form of e-publications (e-journals, e-books, blogs…) provided by electronic, "immaterial" means. Publications on a physical medium such as a CD-ROM are already covered in the same article of the Code du Patrimoine, having been included in legal deposit in the 1992 law.

Other articles cover the responsibility of producers, notably regarding the provision of technical information necessary for the collection and conservation of material and the practicalities of the collection[8], and the conditions of access[9], points which are discussed in greater detail below. The law also specifies that the exact details regarding its actual enforcement will be fixed in a decree (or "décret"; which is the usual process for the practical implementation of legislation in France). It is important to note that, at time of writing, this decree is still in process of validation and is yet to be published; the implementation of digital legal deposit as it is presented here, although put into practice by BnF for several years, must therefore be considered as still experimental. Certain developments and details regarding its

---

[5] *Loi n°2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information.* http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006054152&dateTexte=20110520 (consulted 20 May 2011)

[6] *Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:EN:NOT

[7] *Code du patrimoine*, article L131-2. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000020905828&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

[8] *Ibid*, article L132-2-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845522&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

[9] *Ibid*, article L132-4. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845526&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

implementation will only be confirmed or clarified once the decree is published. References are sometimes made in this article to the most recent drafts of the decree, but these cannot be considered as definitive and the possibilities discussed in relation to the decree remain hypothetical.

b) Decree on legal deposit (1993, modified 2006)

There is another text that is relevant to digital legal deposit. The decree that implemented the legal deposit law of 1992 was modified in 2006 to allow the BnF to propose to publishers that they provide, in place of a physical document, a digital file identical to it, with the manner of the deposit to be agreed between the BnF and the publisher[10]. As discussed in more detail in section III. (f) below, this has so far only been used in the case of very large publicity posters, which are unwieldy and difficult to manage and to consult in their physical format, and are now deposed as PDF files; there has also been e-deposit experiments conducted with one of the major French regional newspapers, *Ouest France*. However the possibility of digital substitution may provide many more other options to be explored in the future; yet it is important to note that this disposition requires that the digital version be exactly identical as to the one distributed in printed form and that it only allows for a replacement of the deposit of a physical document. This option could not for instance be used to collect both the electronic and paper versions of a novel: it does require that a radical choice be made by the Library, to abandon the printed version.

c) Decree founding the BnF (1994)

As already noted, the decree creating the new Bibliothèque nationale de France[11] places the role of legal deposit as central to the missions of the library. In fact, the main legislative base for this mission is still the 1992 legal deposit law, as integrated into the Code du Patrimoine, and the related 1993 decree, as described in the two previous sections. However this other decree establishing the new BnF highlights and reinforces the status of legal deposit collections as part of the national heritage, which has implications especially for questions of long term conservation.

d) Code général de la propriété des personnes publiques, Code de la propriété intellectuelle and Loi relative à l'informatique, aux fichiers et aux libertés

While not directly concerned with legal deposit, three other pieces of legislation are important to mention for the practical application and implementation of legal deposit. The Code general de la propriété des personnes publiques (Code of public property)[12], the Code de la propriété intellectuelle (Intellectual Property Code)[13] are both wide-ranging collections

[10] *Décret n°93-1429 du 31 décembre 1993 relatif au dépôt légal*, article 9.
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082758&dateTexte=20110520
(consulted 20 May 2011)
[11] *Décret n°94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France*.
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082797&dateTexte=20110520
(consulted 20 May 2011)
[12] *Code général de la propriété des personnes publiques*, article L2112-1.
http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006070299&dateTexte=20110520
(consulted 20 May 2011)
[13] *Code de la propriété intellectuelle*.
http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006069414&dateTexte=20110520
(consulted 20 May 2011)

of legislation, and as we will see later in this discussion, the collection, preservation and consultation of material under digital legal deposit is framed by several provisions in these two codes. Finally, another law, the 1978 Loi relative à l'informatique, aux fichiers et aux libertés (Law on information technology, files and freedoms)[14], has a strong bearing on the provisions for access to and usage of digital legal deposit collections as it imposes strict restrictions as to the protection of personal data that may be included in such collections.

e) Summary of the legal possibilities for digital legal deposit

These legal texts therefore leave three possible mechanisms for the collection of electronic material under legal deposit:
Under the Code du Patrimoine, modified by the DADVSI law of 2006:

- automatic collection of material via the Internet (by means of harvesting),
- deposit of digital files by the publisher (by means of "e-deposit").

Under the decree of 1993, modified in 2006:

- deposit of strictly identical digital files as a replacement for paper deposit.

Both for economic and heritage reasons, BnF has so far prioritised the automatic collection of Internet material, and this article examines in particular this aspect of digital legal deposit. However the full range of possibilities will be discussed as offering other approaches to be explored in the future.

The following sections discuss, in order, the four aims of legal deposit as defined by the Code du Patrimoine: the collection of material (section III), its preservation (section IV), the creation of national bibliographies (section V) and the consultation of the collections (section VI). In each case, the legal restraints and possibilities are discussed in relation with the practical measures already in place, and those which may be imagined for the future.

III. Means of acquisition of electronic materials by legal deposit

The legal deposit of electronic publications, while it is in the tradition of earlier forms of legal deposit, creates challenges specific to the nature of the material. As shown in the previous section, the legal texts governing legal deposit allow for a wide range of electronic materials to be included; however the nature of such electronic materials means that two guiding principles underlying the French approach to legal deposit – the idea of publications *being made available on the French territory*, and the *exhaustive nature* of legal deposit – must be reinterpreted.

a) Scope of material subject to digital legal deposit in France

According to the Code du Patrimoine, everything that is published on the Internet in France is subject to legal deposit. This raises the question of how to define the "French Internet"; by definition all information accessible on the web is available in France, and therefore a definition based on that applied to printed books, where imported material is collected, would rapidly become unworkable. The definition which should be given in the forthcoming decree,

[14] *Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.* http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20110520 (consulted 20 May 2011)

and which is already applied in practice by the BnF, is based on the idea of a link to the French territory. Three criteria are used to judge if a publication is in the national scope of electronic legal deposit:

- if it is made available on the French national web Top Level Domain (TLD), .fr, or any other domain name registered within a domain name registry based in France (for instance domains with the .com extension which are registered in France);
- if the producer of the website (or other document) is a person resident in France, or a company based in France;
- if the website is produced in France (this latter criterion being subject to wider interpretations than the former, it also allows for some flexibility).

It is important to note that, in current practices and given both the scale at which BnF operates and the limitation of the resources available, such conditions aren't systematically checked by the Library before harvesting websites. This general definition of the national scope is however taken into account to define the general policy and technical settings of web collections, as the main entry point to national, bulk, domain crawls are currently seeds or addresses of websites registered under the .fr extension. The listed conditions may also be opposed, during or after harvesting, on the basis of individual claims by producers for instance (see section (c), below).

As discussed below, this represents a significant number of domain names, and a huge volume of data. The question may be asked, however, if a national division of the Internet has much sense, as hyperlinks do not respect national borders, and the Internet is by its very nature international. It remains the case that national legal deposit legislation is a powerful means of ensuring large-scale preservation of the Internet, by allowing legal means of copying and preserving content, mobilising the resources of national libraries and archives and placing Internet archiving in the context of the preservation of cultural heritage. The division by countries does however pose the question of international collaboration and interoperability between collections, discussed below (see section (f)).

b) The institutions responsible for digital legal deposit

The Code du Patrimoine distributes the responsibility for legal deposit between three cultural institutions: the Bibliothèque nationale de France, the Institut national de l'audiovisuel (INA, the French national broadcasting Archive) and the Centre national du cinéma et de l'image animée (CNC, in charge of preserving motion pictures)[15]. Regarding specifically digital legal deposit, the forthcoming decree should define the division of responsibility between the BnF and INA. In the meantime, an *ad hoc* division has been agreed between the two institutions, following the logic of the continuity of their respective mandates and collections: INA collects Internet publications relating to television and radio broadcasting in France, and the BnF collects all other material. This division should be fixed more precisely in the forthcoming decree. In this article, the focus on the practical aspects of the collect of Internet materials is based on the experience at the BnF; INA has a different approach based on much

[15] *Code du patrimoine*, article L132-3. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000020967935&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

more frequent crawls of a smaller number of sites, with a strong, complementary focus on stream media[16].

While the legal responsibility for the collection and its display lies with the BnF and INA, other institutions and organisations may be involved in the process, particularly where there is a selection of material to be collected. The BnF has already put in place experimental cooperation with the 25 French regional libraries charged with receiving legal deposit from printers (known as Bibliothèques du dépôt légal imprimeur, or BDLI); these libraries have been involved in selecting sites from their respective regions to be archived during national or regional election campaigns[17]. Researchers and specialists from a variety of organisations (universities, associations…) have also been involved in selecting sites for other thematic or event-based projects and datasets such as web activism, online literature or sustainable development and the green web. Such possibilities should be further explored, although there are implications in terms of providing access to the collections, which are discussed below.

There are in France other initiatives in the area of web archiving, outwith the context of legal deposit legislation. Some researchers and universities are actively engaged in research and development projects regarding the web, and this may involve archiving web material. Most active is the Internet Memory Foundation (previously known as the European Archive Foundation), a not-for-profit foundation that aims to preserve the Internet[18] and, more recently, the Medialab at Sciences Po[19] in the social sciences area. However, only INA and BnF can benefit from the specific dispositions attached to the legal deposit legislation, in particular the possibility to harvest websites without asking permission to the publishers.

While in France, which has a well-know tradition of administrative and cultural centralism, the question of the distribution of tasks between heritage institutions is mainly discussed in relation to the respective mandates of the BnF and INA, which are both located in Paris, other countries have to clarify the distribution of tasks from a different perspective. A frequent situation involves questioning the division between the National Library and the National Archives (this is the case in the UK, for instance). Another one, in the case of federal administrations, requires envisioning more largely distributed and cooperative organisation schemes, such as networks of regional, specialised libraries (as in the case of Switzerland or Germany). Regardless of specific legislations, other forms of networks may develop as in the case of the United States. There, under the umbrella of the National Digital Infrastructure Preservation Program (NDIIPP)[20] led by the Library of Congress, one finds a variety of institutions actively engaged in web archiving such as the non-for-profit foundation Internet Archive, the California Digital Library or the University of North Texas.

c) Questions of scale and completeness

Legal deposit, as defined in French law and by its tradition, has previously aimed at an ideal of exhaustiveness: the resulting collections should contain everything published or imported

[16] For further details see the INA website (in French): http://www.ina-entreprise.com/entreprise/activites/depot-legal-radio-tele/depot-legal-web.html (consulted 20 May 2011)
[17] Huchet, Bernard ; Illien, Gildas ; Oury, Clément. "Le temps des moissons. Le dépôt légal du Web : vers la construction d'un patrimoine cooperative". In : *Revue de l'Association des Bibliothécaires de France*, 2010, n° 52, pp. 28-31
[18] Internet Memory Foundation. http://internetmemory.org/en/ (consulted 20 May 2011)
[19] Sciences Po, Médialab. http://www.medialab.sciences-po.fr/index.php?page=home (consulted 20 May 2011)
[20] NDIIPP website: http://www.digitalpreservation.gov/ (consulted 20 May 2011)

in France, within the criteria defined above. However the extension of legal deposit to digital material means that this ideal must be questioned. The definition of electronic material in the Code du Patrimoine is, as we have seen, formulated so as to be independent of any precise *format* (e-book, web…), rather it places the emphasis on *content* that is communicated by electronic means. This widens the field of legal deposit to include everything published on the web that meets the criteria described above, regarding territoriality and the public nature of any communication. This creates a difficulty, as the very nature of the web seems opposed to any idea of an exhaustive collection.

On one level, this problem comes from the sheer amount of information available online. In April 2011, the number of domain names registered in .fr was around two millions, and to this must be added sites within the remit of French legal deposit registered with other TLDs, notably .com, .org and .net; AFNIC, the body in charge of administrating the .fr TLD, estimates that this represents only a third of the "French internet", using a definition very similar to that applied by the laws on legal deposit[21]. While the national domain crawl performed by the BnF in 2010 showed that a large proportion of these domain names had little or no content, some large sites contain many millions of individual files.

Distribution of domains in terms of number of URLs collected by domain, BnF Domain Crawl 2010

Number of URLs collected	Number of domains
=<10	976,948
11-100	580,362
101-1000	320,620
1001-10000	85,471
10001-50000	23,630
50001-100000	352
>=100001	230

Rather than consisting of individual, separate publications, the web is an information space with shifting boundaries, where it is difficult to define distinct and stable "items" or "units" comparable to a book or an issue of a periodical. A website may contain multiple pages, images, video or audio files, documents in the form of PDF or Word documents, applications… In addition, the nature of the web lies in the use of links within and between sites, so that much of the information takes its meaning and significance from its place within a complex network of interconnecting links. To add to this complexity, there is a constant flow of information, as sites are updated with a frequency that varies between and within sites. All of this means that to be truly exhaustive, it would be necessary to collect everything all the time; the technology of web crawling and the storage space involved mean that this is

[21] AFNIC. *French Domain Name Industry Report 2010*, pp. 20-22. http://www.afnic.fr/data/actu/public/2010/afnic-french-domain-name-report-2010.pdf (consulted 20 May 2011)

simply impossible. The collections created would also be huge and unmanageable, both for librarians and end users.

Faced with this impossibility, the only response is to abandon the ideal of exhaustiveness and accept that the legal deposit of the web will collect only a part of what is available. As regards the updating of online material, the forthcoming decree should recognise this problem, in specifying that sites should be collected "at least once a year". However even then, the mass of material means that an exhaustive collection even once a year remains infeasible. There are then two approaches that can be used: *selection* and *sampling*. The former option involves a prior selection of sites to be collected, usually on the basis of a judgement of the quality or the scientific or aesthetic value of the site; it could thus be decided that sites publishing scientific research, government or official publications or literary or artistic works are of greater worth and should therefore be the focus of the collection. This approach is in many ways similar to the acquisition of books chosen by a librarian, with a logic of selecting items that will enrich the research collections. The alternative approach, sampling, is closer to the idea of legal deposit: sites are collected without a prior judgement being made of their "value" or of their potential interest to current or future researchers. Rather the aim is to preserve a representative sample of the national born digital output, which should capture as far as is possible the "character" of the national web at a given time.

Each approach has its limitations: selection requires the definition of criteria, and an investment of time by curators, researchers and others, with the possibility that the sites selected today will not be those considered most important by users in the future. Sampling on the other hand means that important sites may be collected only partially or not at all, while it may be argued that much of what is collected will be of no interest to researchers as the content might be seen as junk (spam, domain squatting sites…) or of low value (personal blogs, advertisement, commercial sites…).

At the BnF the decision has been taken to combine both approaches and to adopt a "mixed model" for web archiving that combines selection and large-scale sampling. The detail of this approach is described in the next section.

d) How online material is collected

To respect the obligations of legal deposit while accepting the realities of the Web, the BnF has thus put in place since 2006 this "mixed model" of web archiving that combines two types of collect: broad or domain crawls, and focused or selective crawls. The former consists of an annual crawl of all the domain names registered in the TLD .fr; this list is provided annually under an agreement with AFNIC. In the future the BnF hopes to be able to include sites registered in other TLDs such as .org, .net and .com also registered in France, which are within the scope of legal deposit and may represent around two-thirds of sites registered in France (see section (c) above). This will require additional agreements directly with the registrars[22]. There is therefore no judgement regarding the quality or value of what is collected; in the tradition of legal deposit everything which falls in the criteria described above is subject to be collected. This annual collect uses technical settings meaning that only a limited amount of data is collected for each domain: in 2010 this was set at 10,000 URLs

[22] For more information on the role of registrars, see AFNIC. Other domain name registries. http://www.afnic.fr/doc/autres-nic_en (consulted 30 May 2011)

(or files) per domain. While this is sufficient to collect the majority of sites in their entirety, large sites and platforms are only partially collected. The idea of this approach is to provide a kind of "snapshot" of the French Web, which while limited both in depth and in temporal coverage, respects the obligation under legal deposit to collect the French web at least once a year (see section (c) above). This allows the collection of a representative sample of French Internet production.

The other, complementary approach, focused crawls, involves collecting sites which are selected by subject librarians at the BnF, and occasionally by other partners (such as regional legal deposit libraries and researchers). While this still falls within the legislative framework of legal deposit, the approach may be considered as similar to the acquisition of books and other resources serving the purposes of a research library: librarians choose sites based on the value and interest of the material, as part of the resources held by the BnF in a given area; the criteria for selecting sites should then be linked to the overall acquisition policy of the collection and acquisition departments of the Library. Sites are selected which may not be collected, or not satisfactorily, in the broad crawl: this may include sites in other TLDs than .fr, and sites or parts of sites that may not be collected because of the size of the domain (large institutional sites, individual blogs…). The focused crawls also permit sites to be collected more frequently than once a year. As of 2011, the BnF has put in place a system of permanent crawling, where sites may be collected annually, bi-annually, monthly, weekly or even daily. This allows, for example, a daily collection of a selection of news sites, to show what stories are on the homepage on a given day, and improves the quality of collection of sites that are updated frequently, or those that do not maintain archives. Depending on the frequency of crawling the depth of collection and the number of files collected per site varies.

Finally, the BnF has put in place an experimental procedure whereby website producers may propose their own site to be collected. This is currently done by means of a message on the pages of the BnF website devoted to digital legal deposit, which gives an email address where nominations of websites may be made. Depending on the results of this experiment, this approach may be developed once the decree is published. (The role of website producers in automatic collection and deposit of files is discussed in the next section, (e).)

Web archiving at the BnF relies on two main pieces of software developed as open-source products in partnership with other institutions: the crawler robot Heritrix[23], which collects the files which constitute the archives, and NetarchiveSuite[24], which allows the planning, programming and monitoring of crawls. It is important to note that such crawls, which involve making copies of the file that make up websites, are only possible because the Code du Patrimoine, as modified under the DADVSI Act, creates an exception from intellectual property legislation. Article L132-4 specifies that copying copyrighted material is authorised "when such reproduction is necessary for the collection, preservation or consultation" of the material[25]. As it is impossible to collect Internet material without making a copy it was necessary to introduce this possibility, and the DADVSI law was specifically intended to deal with such problems arising from incompatibilities between new technologies and the existing

[23] Internet Archive. *Heritrix*. http://crawler.archive.org/ (consulted 20 May 2011)
[24] Netarchive.dk. *NetarchiveSuite*. http://netarchive.dk/suite/Welcome (consulted 20 May 2011)
[25] *Code du patrimoine*, article L132-4. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845526&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

legislation in various areas. (Other aspects of this article are discussed in section VI (b), below.)[26]

e) Obligations of producers and publishers

This article also means there is no need to seek permissions from website producers and copyright holders; such a system of permission-based web archiving exists in many other countries (such as the United Kingdom[27] and the United States[28]), but this makes large-scale domain crawling impossible, as identifying and contacting the owners of millions of domains would be unfeasible. The non-profit foundation Internet Archive, which lacks a legislative support for its world wide archiving task which started as early as 1996, takes an "opt-out" approach, by which material is removed from access on their archive if a website producer or copyright holder complains[29]. Once again, this exception from intellectual property rights only applies within the strict legal conditions governing legal deposit in general, and particularly the controls placed on access to archived material, as discussed below in section VI.

Another provision introduced into the Code du Patrimoine results from the change, regarding digital legal deposit, in the relationship between the producer/publisher and the depository institution: unlike in traditional legal deposit, it is the BnF that collects websites, rather than receiving deposits from the publishers. However the law makes it clear that producers have a responsibility to facilitate the collection of their material if required:

> [The depository institutions] may proceed themselves with this collection using automated procedures, or may determine the modalities in agreement with [the producers]. The use of a code or a restriction on access by these persons cannot create an obstacle to the collection.[30]

The decree may specify further that producers are obliged to provide all passwords or other means necessary to access documents; this may apply not only to sections of websites protected by passwords, but also to files such as audiovisual contents, priced journals or e-books that may be protected by digital rights management (DRM) technology. DRM may limit both the collection of a site but also its long term preservation and the decree should specify therefore that publishers must provide all necessary information and means of access in both cases (see below, section IV). As regards passwords, it is important to note that the basic definition of electronic material subject to digital legal deposit (quoted in section II (a), above) specifies that it must be "communicated to the public", and therefore excludes any material on the Internet which may be considered as private correspondence. Thus, where passwords are put in place to protect private material – notably in the case of private areas of social networks – these areas are outwith the scope of legal deposit. However published

[26] For more information on web archiving at the BnF, see Bleicher, Ariel. "A Memory of Webs past". In: IEEE Spectrum, Mars 2011. http://spectrum.ieee.org/telecom/internet/a-memory-of-webs-past/0 (consulted 30 May 2011)
[27] See UK Web Archive. Legislative Status. http://www.webarchive.org.uk/ukwa/info/about#legislative_status (consulted 30 May 2011); Department for Culture, Media and Sport. Legal Deposit. http://www.culture.gov.uk/what_we_do/libraries/3409.aspx (consulted 30 May 2011)
[28] Library of Congress. Web Archiving. http://www.loc.gov/webarchiving/index.html (consulted 30 May 2011)
[29] Internet Archive. Terms of Use. http://www.archive.org/about/terms.php (consulted 20 May 2011)
[30] *Code du patrimoine*, article L132-2-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845522&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

material for which a fee is charged, and therefore which is accessible by password (or other means such as IP recognition), is likely to come into the scope of digital legal deposit as it is in print legal deposit. Publishers will therefore be obliged to provide all necessary help to ensure their collection by the Library.

This introduces the idea of how to proceed with the digital legal deposit of material such as the "deep web", databases and e-books, which remain open questions and are discussed in the next section.

f) Open questions/challenges

Unlike other institutions, such as the Royal Library of The Nederland, who started early with e-deposit based on agreements with publishers and begun its web archiving program only several years later, the BnF has so far concentrated its resources in digital legal deposit on web archiving, principally for economic and practical reasons: large amounts of publicly available online material that would otherwise have been lost for ever have been collected using automatic web harvesting. As early as 2004 (when BnF launched its first experimental domain crawl, in partnership with the Internet Archive), the Library felt it was urgent to start collecting at scale as the web changes extremely fast and lots of data of heritage value disappears everyday. However some material has remained difficult or impossible to collect by these means, either because the material is not online, or because technical or commercial barriers limit access. The legislation, as noted above, applies to all published electronic material and obliges producers to cooperate if necessary, which opens the way to other approaches such as deposit of files by producers.

This approach has not yet been implemented at the BnF, partly because the obligations of the producers and the BnF need to be defined more precisely in the decree, and it is therefore preferable to wait for this legal backing. However preparatory work is being done regarding the procedures, both technical and organisational, that will have to be put in place. One area where deposit of files has been experimented is newspapers, which pose specific problems in terms of storage and conservation. Between 2005 and 2009, experiments were conducted with *Ouest France*, a regional newspaper title with a significant distribution in the western part of the country, to organise the e-deposit of the many (currently: 47) local editions of this newspaper, so that the library would only collect and conserve one main edition of reference in printed form, hence saving a lot of manipulations both to the BnF and the publisher, along with significant storage space in the stacks. This experience had to be interrupted for lack of proper resources and because it appeared impossible, in the current state of the art, to expand this workflow to other newspaper titles. It is probable that this will be approached rather by collecting the online versions available on the Web, though this in itself poses other technical and legal issues, discussed below.

Some of these problems can be addressed using web archiving, but there are many issues to be dealt with. There are today many kinds of material that are difficult to collect: rich media (videos, streaming, Flash, javascript…), deep web (databases…), subscription and password-protected contents… Some of the problems are largely technical in nature, and the BnF hopes to benefit from, and contribute to, international efforts to develop tools and skills to improve the collection of rich media, for example. Others are however a combination of technical barriers and legal or organisational questions. In the case of password-protected resources or those requiring IP authentification, the Heritrix robot is able to bypass login pages when

programmed with the access details; it is of course necessary to contact the producers to obtain these details. As explained above, the Code du Patrimoine, supported by the forthcoming decree, obliges producers to cooperate with the BnF and to provide all necessary information for materials falling into the scope of legal deposit, whether they are free of charge or payment-based. As this is a new legal obligation, it is yet to be seen how easy it will be to encourage producers and publishers to collaborate with the BnF in this area; also it will require a lot of work and resources from the Library to maintain and follow-up contacts with publishers while the Library is confronted with severe budget cuts. In certain cases, it will also be necessary to combine this further technical developments, for instance in a case where the password gives access to a Flash-based reader that prevents Heritrix from collecting the content. This work will only start seriously once the decree is published, and will imply much organisational change to deal with this new manner of collecting material. However it should allow the BnF to collect much material, and notably e-books, online journals and other valuable electronic resources that are not currently collected.

For some publications however, it is likely that only deposit of files by publishers will allow the BnF to respect its obligations under legal deposit. For the moment, apart from the experiments regarding the press, the only systematic digital legal deposit performed by publishers concerns large format posters, which, as noted above, are collected in digital form instead of in paper format, under the 1993 decree (modified in 2006) controlling legal deposit. However with the increasing interest and commercial viability of e-books, it remains an open question how best to manage the collection of these publications. Within the structure of the BnF this question is being asked both by the teams responsible for digital legal deposit and the legal deposit of printed books, as well as subject librarians seeking to acquire e-books in their subject areas. The workflow put in place to collect such publications will also have an impact on other areas, including the cataloguing and the preservation of these works. The BnF is therefore examining the options to best respond to this challenge.

A related question regards what may be termed digital gifts or donations. In France, legal deposit was always complemented with acquisitions or gifts, the addition of these different means of collecting contributing to build the heritage and research collection at large. It is an interesting question to see whether similar combinations can be envisaged as to born digital resources: we know of digital acquisitions of course (priced electronic resources), but we can think of digital gifts or “manuscripts” as well (authors, artists…). How will these collections legally and technically interact with the digital legal deposit? Again these questions will need to be discussed in the context of a global solution for electronic publications beyond that already existing via web archiving.

Finally, an avenue for further exploration in the collection of digital material is that of international collaboration. Many initiatives are already in place, particularly in the context of the International Internet Preservation Consortium (IIPC)[31]. As previously noted, these can concern new technologies to improve the quality of web harvesting; already, both Heritrix and NetarchiveSuite (initially developed by one institution, Internet Archive and NetArchive.dk, respectively), or other popular software such as the Web Curator Tool (jointly maintained by the British Library and the National Library of New Zealand), are open-source tools which are developed by the international community. More recently, discussion has turned to how international cooperation may play a role in the selection and

[31] International Internet Preservation Consortium. http://www.netpreserve.org/about/index.php (consulted 20 May 2011)

collection of material, and to some extent address the limitations of a "national" division of the Internet, mentioned above. Thus in the case of events of an international interest each institution could collect web material from its own country. Experiments around this idea have already taken place: IIPC projects relating to the 2009 European Elections and the 2010 Winter Olympics, with a view to collecting the forthcoming 2012 Summer Olympics; and also ad hoc collections responding to urgent situations: the earthquakes in Haïti in 2010 and Japan in 2011, or the political events in North Africa known as *Jasmin Revolution* in 2011. Internet Archive performed crawls based on propositions from different institutions[32], and some institutions (such as the BnF) additionally performed their own crawls. Future work should help to put in place procedures and best practice for such federated collections, however questions remain over interoperability and how to make collections in different countries "talk to" each other, given the legal restrictions on access that exist in many countries' legislations. The access aspects of international cooperation are thus discussed below, in section V.

IV. Conservation of material obtained under electronic legal deposit

a) The legal obligation to preserve heritage collections

As discussed at the start of this article, the aim of legal deposit is to create a permanent record of the cultural output of France. The idea of conservation is therefore at the heart of legal deposit: the collections created by legal deposit must be preserved without restriction of time. This responsibility, implied by the decree creating the BnF and the Code du Patrimoine, ultimately draws its legal force from the Code général de la propriété des personnes publiques[33]. Article L2112-1 of this code specifies that one copy of each document collected under legal deposit (those listed in article L131-2 of the Code du Partimoine) must be considered as part of the "Domaine public mobilier", or items belonging to the public domain, and therefore "inaliénable et imprescriptible"[34]. This fundamental obligation places a special importance on the role of preservation techniques, where, as with the collection of material, the nature of electronic legal deposit poses both technical and legal challenges different from those previously encountered with other media.

b) Technical approaches to digital preservation: making copies for conservation

From the technical point of view, systems are being put in place at the BnF to ensure the long-term preservation of digital legal deposit collections. Similar systems are currently being built in a growing number of national libraries, such as the Library of Congress, the national libraries of New Zealand and Australia. A major difference in the preservation of digital material is the need to be able to make copies, either identical copies, or modified copies to allow for changes in format, etc.; indeed, digital preservation is impossible without the ability to copy from one support to another. The DADVSI law of 2006 created an exception to

[32] Internet Archive Global Events. http://www.archive-it.org/public/partner.html?id=89 (consulted 20 May 2011)
[33] *Code général de la propriété des personnes publiques*, article L2112-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006361198&cidTexte=LEGITEXT000006070299&dateTexte=20110520 (consulted 20 May 2011)
[34] *Ibid*, article L3111-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006361404&cidTexte=LEGITEXT000006070299&dateTexte=20110520 (consulted 20 May 2011)

intellectual property laws, now in the Code du Patrimoine as Article L132-4[35], which allows for the copying of copyrighted material where this is necessary for the collection, as discussed above, but also for the preservation of material..

Currently, the web material collected (in the form of ARC or WARC files[36]) is stored in the first instance on the servers used for access to the collections. A security copy is also made, with checks and backups in place to prevent the deterioration of the files or the storage media, such as checksums, periodic copies and replacement of hard disks/tapes. This security copy for bit-stream conservation is a first step in preservation, however for long-term preservation the BnF will integrate its web archives into the centralised digital preservation system known as SPAR (Scalable Preservation and Archiving Repository)[37]. This repository complies with the principles of the Open Archival Information System (OAIS) model[38]. In addition to checks on the integrity of the data, SPAR will allow an analysis of the formats used, which in turn will permit long-term preservation strategies to combat format obsolescence, based on migration and emulation[39]. The article of the Code du Patrimoine already cited allows this type of manipulation of the data necessary to conserve legal deposit collections. As noted above, the forthcoming decree should also reiterate the obligation for producers to provide all technical details necessary for the conservation of legal deposit material; this will apply notably in the case of digital rights management (DRM) technology which could limit the reproduction or modification of files as envisaged in SPAR.

c) Requests for destruction or modification of material, from an individual or the publisher

The BnF also has to consider the possibility that individuals may request that information held in the web archives be modified or destroyed. This is particularly in view of the 1978 Law on Information Technology and Freedoms, which allows individuals to correct or delete information regarding them published on a website, or held by a third party[40]. Requests may also be received from the author or publisher of material seeking to remove or modify material. However, the legal obligation to preserve legal deposit collections, as described in section (a) above, overrides any other rights or requests to destroy material, and the same protection should be applied to digital collections as to books and other physical collections, which cannot be destroyed or modified. This legal safeguard is vital for preserving the integrity of heritage collections, and at the heart of legal deposit.

---

[35] *Code du patrimoine*, article L132-4. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845526&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

[36] Internet Archive. *Arc File Format.* http://www.archive.org/web/researcher/ArcFileFormat.php (consulted 20 May 2011)

[37] BnF. *Preservation of digital material: the SPAR project.* http://www.bnf.fr/en/professionals/preservation_spar/s.preservation_SPAR_presentation.html (consulted 20 May 2011)

[38] ISO. *Open archival information system: Reference model (ISO 14721:2003).* http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683 (consulted 30 May 2011)

[39] Bermes, Emmanuelle; Fauduet, Louise; Peyrard, Sébastien. A data first approach to digital preservation: the SPAR project. In: *World Library and Information Congress: 76th Ifla General Conference And Assembly (IFLA 76)*, 10-15 August 2010, Gothenburg, Sweden. http://www.ifla.org/files/hq/papers/ifla76/157-bermes-en.pdf (consulted 20 May 2011)

[40] *Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.* http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20110520 (consulted 20 May 2011)

This is an area that should be clarified in the forthcoming decree. The body responsible for issues of privacy and data protection in France is the Commission Nationale de l'Informatique et des Libertés, or CNIL (National Commission for Information Technology and Freedoms). It has recently published a recommendation regarding the decree which recognises that the right to modify or delete material shall not apply to legal deposit collections. It is therefore envisaged that such requests will be dealt with by limiting or removing public access to the material in question, rather than destroying it; the issues relating to this possibility are discussed in section VI below. Finally, it should be noted that a court decision may result in an order to destroy or remove material held in the web archives; as with printed material, the BnF would be obliged to respect such a judgement, but such cases are extremely rare.

d) Open questions/challenges for the future

A major challenge for the future, already envisaged with the creation of SPAR, is the risk management of web archive collections. While the long-term preservation strategies outlined above are possible, any intervention in the collections has costs associated in terms of both human and machine time spent analysing and processing collections and increased storage space. It may not therefore be possible to apply the highest quality preservation strategies to all material in the archives – instead a system of risk management may be put in place, in which certain formats, which for instance may be considered as presenting a particular risk of obsolescence, could be given higher quality treatment than others. This may be compared with the approach used for paper collections, where works whose physical medium is considered fragile are given a different treatment and, for instance, stored in special stacks or communicated to the public under special conditions. The overriding legal obligation to preserve digital legal deposit collections does not change but, as with the movement from exhaustiveness to sampling, given the amount of data collected under legal deposit, a pragmatic approach might need to be taken in order to manage limited preservation resources in a sustainable way.

V. Description of legal deposit collections and national bibliography – what obligations exist for cataloguing material obtained under digital legal deposit?

a) The legal requirement of description

The nature of electronic material, and particularly the Web, also requires a new approach to the obligation to create a national bibliography of the French national production, which is given in the Code du Patrimoine as one of the aims of legal deposit[41]. This obligation is put into place by the publication of the French National Bibliography by the BnF[42], and catalogues made available by INA and the CNC. While electronic publications which respect the form of existing publications, such as e-books and online journals could in theory continue to be treated in the same way, the nature of information published on the Web means that web archives require a different treatment.

[41] *Code du patrimoine*, article L131-1, (b). http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845515&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)
[42] BnF. *French National Bibliography Online*. http://www.bnf.fr/en/professionals/bibliographic_products_and_services/a.french_national_bibliography_online.html (consulted 20 May 2011)

As described above, the Web is different not only in the scale of the information that it produces, but in the nature and granularity of it: rather than being individual, separate publications the Web can be seen as a constant flow of interlinked data. Applying a traditional bibliography approach is therefore impossible at large scale, due to both the difficulty of identifying "items" to be catalogued (websites, domains, hosts, directories of websites, pages, files…) and the sheer amount of material (the 2010 French domain crawl covered 1.6 million domains, and collected some 800 million files).

The forthcoming decree, in recognition of this difference, should allow for automatic indexing to take the place of traditional bibliography as the means of ensuring access to the collections. This is in line with the measures already put in place by the BnF, which has chosen not to catalogue manually the material collected by Web harvesting, but has instead put in place indexing. This approach is based on the idea that the Web archives should follow the same logic as the live Web, where sites are interconnected rather than viewed as separate, stable units. Other institutions have however taken different approaches, especially when their web archiving program is by law or policy restricted to rather selective datasets; for example the British Library and the Library of Congress both catalogue individual sites. This is in part possible due to the fact that they do not currently perform large-scale crawls; it is doubtful whether this approach could be applied to a domain crawl; see section III (e) above, notes 24 and 25.

b) The technical means currently in place

An experimental interface allowing access to the BnF web archive collections has been available since 2008. This is based on the open-source Wayback Machine, originally developed by Internet Archive[43]. However for the moment there is no comprehensive full-text indexing of the web archives; experimental full-text indexing has been applied to around 10% of the collection, but due to prioritisation of resources on the collection of material, it has not been possible to devote development time to the creation of full-text indexing. This will be a major project in the near future, as to fully exploit the nature of the web archives a functional "search engine", with the addition of the temporal element, is vital. Several institutions, such as INA, the British Library or the National and University of Iceland have been more successful in implementing full text search engines for their web archives. Although they still have to deal with ranking and relevancy issues, they demonstrated that such a project is perfectly doable but requires a fair amount of computing resources.

At the BnF, the principal means of access currently is therefore a search by URL, similar in principle to the search functions made available on the Internet Archive's website www.archive.org. This level of indexing means that the entire collection contained in the web archive is therefore available for access, as required by the legislation. However there are clear limitations to this means of access: it is necessary to know the URL of the site you are looking for, and there is no continuity in the archives for a site that changes its URL over time. For example, the political party currently in government in France, the UMP, has changed its website from http://www.u-m-p.org/ to http://www.lemouvementpopulaire.fr/ and the shift from one version to the next isn't explicit while browsing the web archive. It is also impossible to perform thematic searches in the archives, let alone to perform more sophisticated data mining analysis (see section VI (e) below). It also means that, at least for the moment, the Web archives are not accessible via the BnF catalogue, but only using their

[43] Internet Archive. *Wayback*. http://archive-access.sourceforge.net/projects/wayback/ (consulted 20 May 2011)

own dedicated interface, secluded from other Library federated search applications. Other institutions, such as the British Library or the National Library of Singapore have done a very interesting job in better integrating their web archive to the rest of their digital collections and it is obviously another key area to investigate in the future, also in order to demonstrate and enhance the value of the web archives to end users.

c) Open questions/challenges for the future

These limitations can be seen to restrict the access to the archives, and even if the legal framework allows indexing (including URL indexing) as a replacement for the obligation to produce a national bibliography, it is clear that to respond to the spirit of this obligation, and to allow researchers to analyse the online cultural production in France, it will be necessary to implement new technological solutions matching the researchers' expectations and emerging data and link mining practises on the life web.

Again, the question of how to put in place international cooperation is relevant. Where collections have been created in the context of an international cooperation, as described above, even where an online access is impossible for legal reasons, as in France, it may be possible to put in place systems allowing users to search across multiple countries' collections, to at least see what material is held in different institutions. This is the subject of discussion within the IIPC consortium; technically, from the point of view of the BnF, the developments opening up the web archives to searching from the catalogue could be a step towards this.

Finally, the nature of electronic resources, and particularly web archives, opens possibilities of creating other means of access, beyond "description" of resources as traditionally practised. A first step in this direction will be the implementation of full-text indexing, however other tools could be put in place allowing much more detailed processing of electronic material by means such as data mining. In Europe, LAWA (Longitudinal Analytics of Web Archive Data)[44] project, supported by the European Commission, is exploring interesting use cases and prototypes to move further in this area. The implementation of such tools which require the manipulation of large datasets in specific software environments is yet largely linked with the legal situation regarding access to these resources, and these questions are therefore discussed more fully at the end of the next section.

VI. Conditions of access to material obtained under electronic legal deposit

a) Providing access vs. protection of the collections

In general terms, there are three models for access to web archives: a dark archive, where material is collected but not made accessible (or at least not until an embargo period has passed); a white or "open" archive, which is entirely open to the public (via the Internet); and a "grey" archive, which provides controlled access under certain conditions. Currently some institutions, such as the National Library of Norway, are obliged to have dark archives, which is usually due to copyright and especially data protection laws in force in their country; in these countries the conditions of any access to archived web material has yet to be agreed. Other countries have put in place open archives, often made possible by the fact that their

[44] LAWA website : http://internetmemory.org/fr/index.php/projects/lawa1 (consulted 20 May 2011)

collection system is permission-based, and the request to make material publicly available is included in the permission request; this is for instance the case in the United Kingdom[45]. As noted above, Internet Archive[46] also maintains an open archive, but with a take-down policy in the case of complaints. The National and University of Iceland is, to our knowledge, the only public heritage institution who chose to take a similar approach.

The French legislation allows for material collected under digital legal deposit to be made available only in a grey archive, with strict restrictions on its access. This creates tension between the obvious need to provide access to the archives, which otherwise would serve no purpose, and the legal restrictions.

One of the missions of the BnF, given in the founding decree cited above, is:

> To ensure access by the greatest possible number to the collections, with the exception of secrets protected by law, under conditions respecting the legislation on intellectual property and compatible with the conservation of the collections.[47]

This shows well the tension between the needs of access and those of conservation, and the need to respect intellectual property rights along with personal data protection. However unlike with paper and other physical media, the use of electronic resources does not pose a risk of damaging them, since, as described above (section IV (b)), the law allows for copies to be made specifically for conservation purposes, and the system being put into place at the BnF will ultimately separate the copies used for access and preservation. It is therefore especially the aspect of intellectual property rights that comes into play. The Code du Patrimoine[48] specifies that electronic material collected under legal deposit may be consulted "on site by duly accredited researchers… on individual workstations the usage of which is exclusively reserved to these researchers". The decree should maintain this limitation, although it may allow access to be provided by a limited number of other libraries sharing with BnF legal deposit responsibility. In such case, access from these regional libraries would still involve to maintain the requirement of individual workstations and accreditation for researchers. These possibilities are discussed further in section (e) below.

Access to the web archives is provided in the research reading rooms of the different sites of the BnF. These reading rooms are reserved for researchers who have a demonstrable need to use these collections; this limitation is in place to protect the physical legal deposit collections, mainly for the needs of preservation discussed above. In the case of web archives, there are other reasons that justify this form of access.

b) Authors' rights and intellectual property

It is important to note once again that the 2006 law that created digital legal deposit, and added these articles to the Code du Patrimoine, was intended to deal specifically with

[45] UK Web Archive. http://www.webarchive.org.uk/ukwa/ (consulted 30 May 2011)
[46] Internet Archive. http://www.archive.org/web/web.php (consulted 30 May 2011)
[47] *Décret n°94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de Fra*nce, article 2. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082797&dateTexte=20110520 (consulted 20 May 2011)
[48] *Code du partimoine*, article L132-4. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845526&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulted 20 May 2011)

questions of authors' rights and intellectual property, in the face of the changing landscape of the "information society". This accounts for the situation, which may appear paradoxical, that the BnF currently collects material which is freely available on the web, and then only provides access to it under these strict conditions. On a simple level, this prevents a kind of "competition" between the archived versions of a site and the live website: if the archives were available directly on the Internet, and indexed by search engines, an out-of-date version of a website could theoretically be placed higher than the current version. Even without this, authors and publishers of websites will prefer people to use the "live" site, as it allows them to track usage, generate income from publicity, and so on. Moreover, if the BnF starts to collect sites requiring payment, this controlled access will become even more important, as of course the Library cannot provide online and freely resources for which the publishers charge a subscription.

c) Right to privacy, data protection and sensitive material

Another practical reason for access limitations comes from the questions of privacy and data protection, discussed above (see section IV). The 1978 Law on Information Technology and Freedoms allows individuals to correct or delete information regarding them published on a website; this is overseen by the CNIL. The existence of digital legal deposit means that older versions of the website, containing the erroneous or otherwise unwanted information, may still be available in the archives of the BnF. As described in section IV (c) above, the CNIL has recently acknowledged in a recommendation that the BnF cannot be required to modify or destroy such material under the 1978 law, however the BnF may decide to put in place additional restrictions on access in such cases.

Other similar cases arise from material found to be defamatory by a court, or material that is illegal. Finally, some material in the archives that may be legal but potentially offensive, for instance pornography, may require special controls on access. (In the final case it should be noted that access to the research library, where the archives are available, it already limited to people aged 18 or over.)

d) Practical implications for the BnF

In these cases, the access limitations imposed by the law provide a first response, as potentially sensitive material is protected *de facto* from access by the general public, and may only be consulted in the context of research or other justified use. The CNIL, in its recommendation on the proposed decree, has found that the measures in place at the BnF requiring researchers to undergo a process of accreditation are satisfactory. However as this may not be sufficient to deal with all sensitive situations that arise other solutions may be imagined.

The system proposed by the BnF, which is yet to be implemented, is to put into place a system of "restricted access", by which some material may be held back from the archives as accessible in the research library, and either put in an inaccessible "dark archive", or be accessible only on special demand and with additional justification required. A similar system already exists for printed and other material, and the same principles are likely to be used in the case of web archives. Thus, the BnF is obliged to restrict access where a court judgement specifically demands it; it is also obliged under the Code du Patrimoine to restrict

access to material defined as "secrets protected by the law"[49], which includes material such as military secrets. Additionally, in very rare cases sensitive or illegal material (such as pornography or publications inciting racial hatred) is not accessible by a simple demand but only via a separate procedure. This may also be applied in the case of a book which has been found in a court judgement to be defamatory, and which is therefore removed from circulation; the copy which has entered the BnF under legal deposit cannot be destroyed as this is prohibited by the status of legal deposit collections, as discussed above (section IV), but the system of controlled access protects, in this case, the person who has been defamed, from a widespread dissemination of the material. In all these cases the application of such restricted access is judged on a case-by-case basis with the aim to take into account people's rights and demands regarding their privacy while respecting the heritage mission of the Library.

The Wayback Machine permits the creation of such a system, which is already functional in other institutions, such as Library and Archive Canada, but has yet to be implemented at the BnF. Furthermore, it remains to be decided on what basis such a reserve would be put into place; apart from cases where there is a legal obligation, described above, it will be necessary for the BnF to put into place a system whereby each demand may be judged on its merits. Equally, the question of whether the restrictions are put in place permanently or for a defined period, and the time period before material may be returned to normal access, would also have to be decided case by case, with perhaps a regular re-evaluation of items to judge whether restricted access is still justified.

Another question in the area of consultation regards the reproduction of material in the archives. Again this is controlled by law on intellectual property, just as with other legal deposit collections, and all reproduction is thus strictly limited – the potential problems with electronic resources being that much greater as it is much easier to make identical copies. Currently, copies are limited to printouts of the screen; screenshots are not permitted, much less the copying of files (images, PDF or other documents, HTML code...). The Code du Patrimoine only permits copying insofar as this is necessary for the collection, conservation and consultation of the material; it therefore allows the copying of files between the storage servers and the clients used for consultation (as with the collect, the idea of "copying" becomes problematic since all use of online electronic resources involves copying of files, even if only temporarily), but does not permit more permanent copies. The forthcoming decree should also give more precision in this area, however it is likely to add that access can only be provided using "interfaces for access, search and treatment provided by the BnF, INA and other organisations". This restriction on the treatment of the data in the archives is important, and introduces one of the challenges regarding possible uses by researchers.

e) Open questions/challenges for the future

Currently there is a double restriction on access to the archives which limits their use: the legal restriction requiring controlled access, and a technical restriction arising from the lack of full-text searching and other tools for accessing and handling the archives, as discussed above (sections V (b) and VI (a)). Currently, access is limited to the research reading rooms at the different sites of the BnF. To improve the service offered to end users, the forthcoming decree might allow controlled access in a limited number of regional libraries (the BDLI),

[49] *Code du Patrimoine*, article L131-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845515&cidTexte=LEGITEXT000006074236&dateTexte=20110530 (consulted 30 May 2011)

already involved in the selection of sites, as described in section III (b) above. There is one BDLI in each region in France, and the publication of the decree, signalling that digital legal deposit is now firmly part of the French legal landscape, will allow the means of access already in place to advance beyond the experimental stage, and would therefore allow the BnF to promote this new service more widely as well. To profit fully from such an opening, new tools and new ways of using the archives will be necessary; however the creation of further tools can only be imagined in the context of the limits imposed by the legislation. Possible use cases have already been suggested, which allow us to imagine at least three areas for possible development: more profound use by researchers, requests for copies of sites by the producers themselves, and requests for copies for use in legal cases.

For researchers, the limit on even using screenshots restricts the ways in which they can use the archives in their work. It may be possible to imagine exceptions for academic use of small parts of an archived site, however it is difficult to see how this may be accommodated within the legislation, which seems to rule out any copying or use of digital legal deposit material outwith the BnF or the other approved institutions.

A more serious question lies in the analysis of the archives themselves by researchers. As noted above, the decree specifies that all access to the archives of the BnF must be via interfaces provided by the library. Researchers working with the web, and therefore with web archives, have need of tools that will not only allow them to search the archives (as the full-text indexing will allow them to do), but of data-mining techniques allowing more creative use of the collections. This could include tools to chart the use of certain terms, trends or names over time and across different websites or link analysis of the connections between sites over time… Different research centres are already working with such tools, however the legal restrictions prevent them making copies of the data held by the BnF to be processed by such tools, or to install software themselves at the BnF to analyse the archives. The BnF must therefore envisage ways to make such tools available to researchers, and in particular to allow approved software to be installed. Doubtless partnerships with universities and research centres will be invaluable in this area, and the BnF has just launched a collaborative project with Medialab at Sciences Po to explore such questions[50].

Another case that has already been encountered, and which will no doubt become more common as the web archives become better known and as more material disappears from the live web, is that of a website producer or author who seeks to recover material which is no longer online and of which they have not kept a copy. This may be a website producer whose site was hosted by a third party which ceases to exist, an individual who wishes to keep a copy of a blog they kept in the past and which has disappeared from the platform online, or a journalist who has written material for an online source that is no longer available. Strictly speaking, the legislation stops any copying of material collected under digital legal deposit except for collection, preservation and on-site consultation (in article L132-4 cited above). However as noted above this is primarily to protect the holders of the intellectual property rights. In this case, where it is the rights-holder who requests the material it should be possible to make an exception. The exact means by which a demand may be made are still to be defined, for example who is permitted to make a request (the descendants of someone who wrote a blog a hundred years earlier?). In particular, it will be necessary in each case to prove the right of the person to the material requested, as it would only be possible to provide copies in cases where the demand comes from an individual who holds the rights to all of the

[50] Sciences Po, Médialab. http://www.medialab.sciences-po.fr/index.php?page=home (consulted 20 May 2011)

material in question; for example the rights to music hosted in a blog may belong to someone else. It is also important to note that the technical procedures for exporting all the files relating to a website, or a part of a website, captured at a specific time are not yet in place. SPAR should helps with this, but it will be necessary to implement other technical measures. Comparisons may be made with the reproduction of paper and other material, which is possible for a fee, but only for out-of-copyright works.

Finally, the BnF has already had several demands for copies of web archives in connection with legal cases, to prove the presence of material online at a given date; this may be relevant in cases of intellectual property, or disputes regarding the conditions of sale in web commerce. The BnF, with its status as a national institution, should be able to play a role as a trusted third party for such material, and the means of collection in place, which associate metadata to each file collected, mean that the presence online of a file can be proved at the exact date and time it was collected. However this presents various problems, and in particular the exact conditions under which an exception to the ban on copying legal deposit material may be made are yet to be established; it will be necessary to examine the question in depth to judge what conditions may be considered sufficient to justify an exception to legal deposit law. If such an exception were found to be possible, the BnF would also need to put in place the technical means necessary to produce an "authenticated" copy of the files in question, with the date of collect and the provenance clearly marked. This use of the web archives is yet to be properly tested in a court of law; however if a precedent is established, it may be imagined that this kind of request will become ever more common in the future. This final example does however provide a further demonstration of the interest in maintaining the ideal of a wide-ranging legal deposit of the French web, as embodied by the broad crawl at the BnF; it is impossible to predict which sites, including apparently uninteresting commercial sites, may become important or "needed" years later. It is therefore necessary to collect as widely as possible, as the value of the digital legal deposit collections will only become apparent in the future.

VII. Conclusion

Digital legal deposit in France is only a few years old, but already it has established itself as part of the core missions of the BnF. Beginning with the first experiments in 2002, the BnF has put in place a system of web archiving involving technical solutions, both hardware and software, but also organisational elements, as this mission requires the expertise of digital curators, IT specialists, subject librarians and legal experts, as well as the strong support of the top management. Today the BnF has a robust, flexible and effective workflow for web archiving, which ensures that material published on the French web finds its place in the heritage collections created under legal deposit.

As the various sections of this article have demonstrated, at each part of the chain of digital legal deposit (collection, conservation, description and access) there remain many questions and challenges for the future. These combine both legal questions, and the technical and organisational solutions that need to be put in place to fully respond to the obligations of digital legal deposit. In the short term, the publication of the decree will be an important step: it will establish clearly the legal basis of digital legal deposit, and will allow the BnF to proceed with important projects, such as improving the collection of payment-based publications and possibly widening access to researchers by exploring the possibilities of access in regional libraries.

These two projects, and others outlined above such as collection of rich media, the collection of e-books and other material not collectable using web archiving, full-text indexing, access from the catalogue, federated searching, differentiated preservation strategies, and the creation of data mining and other tools to fully exploit the archives – all of these will require significant technical work and resources to be put into place, and the BnF will have to choose its priorities in the coming years.

One important factor as digital legal deposit matures will be the view that researchers and others have of these collections. Many possible uses can be imagined and have already started to occur, but the legal situation in many of these cases is yet to be firmly established. However the role that the web archives and digital legal deposit collections may have will determine to a large extent the priorities: where there is a clear need for new tools, usage or services, resources will have to be dedicated in those areas.

As a founder of the IIPC consortium in 2003 and a long standing member of its steering committee, the BnF has long been active in the international community in relation to these and many other questions. International cooperation, both technical and organisational, will continue to be central to the development of its digital legal deposit. International collection building, expert discussion, benchmarking, standardisation along with exchanges of skills, staff and best practises have been invaluable assets and created unique opportunities to achieve BnF's web archiving program in the past ten years. As archiving the web is, by nature, a world wide task, international cooperation will remain central for the exploration of the solutions to the many challenges that remain ahead of us.

Bibliography

Legal texts governing digital legal deposit in France (in French)

Code du patrimoine [Cultural heritage code]: title 3, legal deposit http://www.legifrance.gouv.fr/affichCode.do?idArticle=LEGIARTI000006845515&idSectionTA=LEGISCTA000006159934&cidTexte=LEGITEXT000006074236&dateTexte=20110517 (consulted 17 May 2011)

Loi n°2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information [Law on Authors' Rights and Related Rights in the Information Society] http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006054152&dateTexte=20110520 (consulted 20 May 2011)

Code de la propriété intellectuelle [Intellectual Property Code] http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006069414&dateTexte=20110520 (in French); consulted 20 May 2011

Code général de la propriété des personnes publiques [Code of public property] http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006070299&dateTexte=20110520 (consulted 20 May 2011)

Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. [Law on information technology, files and freedoms]

http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20110520 (consulted 20 May 2011)

Décret n°94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France [Decree creating the Bibliothèque nationale de France]
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082797&dateTexte=20110520 (consulted 20 May 2011)

Décret n°93-1429 du 31 décembre 1993 relatif au dépôt légal [Decree regarding legal deposit]
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082758&dateTexte=20110520 (consulted 20 May 2011)

Selected legal texts in other countries

United States:

Copyright Law of the United States of America and Related Laws Contained in Title 17 of the United States Code; Chapter 4: Copyright Notice, Deposit, and Registration.
http://www.copyright.gov/title17/92chap4.html (consulted 20 May 2011)

United Kingdom:

Legal Deposit Libraries Act 2003
http://www.legislation.gov.uk/ukpga/2003/28/contents (consulted 30 May 2011)

Institutions responsible for legal deposit in France

BnF:
http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html (consulted 20 May 2011)

INA: http://www.ina-entreprise.com/entreprise/activites/depot-legal-radio-tele/depot-legal-web.html (consulted 20 May 2011)

Other non-legal deposit institutions in France

Internet Memory Foundation. http://internetmemory.org/en/ (consulted 20 May 2011)

Sciences Po, Médialab. http://www.medialab.sciences-po.fr/index.php?page=home (consulted 20 May 2011)

Other sources

AFNIC. *French Domain Name Industry Report 2010*.
http://www.afnic.fr/data/actu/public/2010/afnic-french-domain-name-report-2010.pdf
(consulted 20 May 2011)

Bermes, Emmanuelle; Fauduet, Louise; Peyrard, Sébastien. A data first approach to digital preservation:
the SPAR project. In: *World Library and Information Congress: 76th Ifla General Conference And Assembly (IFLA 76), 10-15 August 2010, Gothenburg, Sweden*.
http://www.ifla.org/files/hq/papers/ifla76/157-bermes-en.pdf (consulted 20 May 2011)

Bleicher, Ariel. "A Memory of Webs past". In: IEEE Spectrum, Mars 2011.
http://spectrum.ieee.org/telecom/internet/a-memory-of-webs-past/0 (consulted 30 May 2011)

BnF. Preservation of digital material: the SPAR project.
http://www.bnf.fr/en/professionals/preservation_spar/s.preservation_SPAR_presentation.html
(consulted 20 May 2011)

Department for Culture, Media and Sport, "Legal Deposit".
http://www.culture.gov.uk/what_we_do/libraries/3409.aspx (consulted 30 May 2011)

Huchet, Bernard ; Illien, Gildas ; Oury, Clément. Le temps des moissons. Le dépôt légal du Web : vers la construction d'un patrimoine cooperative. In: Revue de l'Association des Bibliothécaires de France, 2010, n° 52, pp. 28-31

International Internet Preservation Consortium. http://www.netpreserve.org/about/index.php
(consulted 20 May 2011

Internet Archive. http://www.archive.org/ (consulted 20 May 2011)

Internet Archive. Arc File Format.
http://www.archive.org/web/researcher/ArcFileFormat.php (consulted 20 May 2011)

Internet Archive. Heritrix. http://crawler.archive.org/ (consulted 20 May 2011)

ISO. *Open archival information system: Reference model (ISO 14721:2003)*.
http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683
(consulted 30 May 2011)

Library of Congress. Web Archiving. http://www.loc.gov/webarchiving/index.html
(consulted 30 May 2011)

Netarchive.dk. NetarchiveSuite. http://netarchive.dk/suite/Welcome (consulted 20 May 2011)

UK Web Archive. http://www.webarchive.org.uk/ukwa/ (consulted 30 May 2011)

UK Web Archive, "Legislative Status".
http://www.webarchive.org.uk/ukwa/info/about#legislative_status (consulted 30 May 2011)