Submitted on: 26.03.2015

Selecting websites in an encyclopaedic national library: a shared collection policy for internet legal deposit at the BnF

English translation of the original paper "La sélection de sites web dans une bibliothèque nationale encyclopédique: une politique documentaire partagée pour le dépôt légal de l'internet à la BnF"

Translated by **Peter Stirling**
Legal deposit department, Bibliothèque nationale de France, Paris, France.
peter.stirling@bnf.fr

Sylvie Bonnel
Coordination service for the Collections Direction, Bibliothèque nationale de France, Paris, France.
sylvie.bonnel@bnf.fr

Clément Oury
Legal deposit department, Bibliothèque nationale de France, Paris, France.
clement.oury@bnf.fr

Abstract

In the space of a few years, the web has become one of the main channels of cultural expression and consumption in French society; online publications have become part of our heritage. This heritage is all the more precious due to its fragile nature. In France, it has been decided to include the preservation of the internet in the centuries-old tradition of legal deposit.

However, adapting this legal and scientific framework to such a vast and wide-ranging space for the circulation of information is far from simple. The BnF defines the scope of its internet collection by a series of concentric restrictions: legal, technical and economic. To ensure that its legal deposit stays representative, the BnF has also adopted an original archiving model, combining broad crawls of the national domain with more focused approaches for sites identified by BnF librarians or by partners.

The BnF has thus been led to apply principles of selection within the context of legal deposit. To this end, each department that participates in the harvesting has developed, by means of successive experiments, its own documentary strategy. The web legal deposit "corresponding officers" have adopted approaches that are not contradictory but rather complementary: selection/sampling, *continuity of collections/exploration of new territories. The BnF must now work on bringing together these different policies, in the context of the renewal of its collection development charter and against a backdrop of budgetary constraints that require the definition of clear priorities.*

Keywords: Collection policy, internet legal deposit

Introduction: Preserving the memory of a new medium

What do the following have in common: Jacques Chirac's campaign website from the 2002 presidential election, the online readers' community Zazieweb, and the "Mauvais Genres" blog dedicated to detective fiction and science fiction? All of them, despite the interest they hold and their popularity when they were online, have today disappeared. Or rather, they would have disappeared had they not been archived by the BnF in accordance with legal deposit.

In the space of a few years, the web has become one of the main channels of cultural expression and consumption in French society. All of the different types of document that libraries have a duty to conserve and make available have experienced, albeit with different timescales, a digital revolution. The web has also created its own forms of expression, which draw on the striking possibilities offered by this kind of network in terms of publishing, of making links between different content and of integrating heterogeneous kinds of documents: as is the case for blogs and social networks.

This exceptional dynamism has however its downside: online content, though accessible throughout the world, is generally hosted on a small number of servers, or even has only a single physical location. A technical problem – a change in the website architecture, server failure… - or a human decision can mean a piece of information disappears permanently. Studies by AFNIC, the organisation that manages the .fr domain, have shown that each year more than 20% of domain names are not renewed. Yesterday's web has already largely disappeared: its principal actors from the late 1990s, such as Geocities or Altavista, have sunk without trace. In ten or fifteen years, how will we be able to remember what today's web was like?

The internet, both in terms of the amount and the variety of the content that it makes available and of the position it now holds in contemporary societies, has become a major part of our heritage. This heritage is all the more precious as it is fragile and highly volatile. Very soon after the birth of the web, the need to preserve the trace of its existence was recognised by a few pioneering institutions: associations and non-profit foundations, such as Internet Archive, and national libraries, such as that of Sweden, started experimenting as early as 1996 with the principles and the methods of web archiving for heritage reasons.

In France, the earliest reflections on the subject date from the end of the 1990s; the earliest practical implementations date from the beginning of the current century. From the beginning it was decided to enshrine this mission within the centuries-old tradition of legal deposit. Created in 1537, this legal principle decrees that each publication produced or distributed in France must be included in the national collections. Over time, it has been adapted to take into account the evolutions in the publishing industry: after printed material, engravings, sound recordings, videos and software have been included among the categories of documents subject to legal deposit. One of the specific features of legal deposit is its nonselective character: all cultural productions are meant to be deposited, whatever the "value" that librarians might accord to them. This principle means that the BnF today has in its holdings works that no other library wanted at the time of their publication. At the same time, this principle is in accordance with the encyclopaedic mission of the institution, reaffirmed in the founding decree of the BnF in 1994.

While "all fields of knowledge", to use the terms of the 1994 decree, are indeed present on the internet, adapting legal deposit to such a huge and wide-ranging space for the circulation of information was by no means simple. How to define the mission of a national institution faced with the inescapably international nature of the web? How to accommodate, at the same time, the requirements of covering publishing output in as complete a manner as possible, as demanded by legal deposit; the aim of preserving by preference the richest and most interesting parts of this output, corresponding to the principles of collection policies; and the technical and economic constraints which make any attempt to build an exhaustive archive of the web impossible?

Faced with these contradictory requirements, the BnF has had to create, over the course of a decade, an original model of collection. This model stems in the first place from the French Code du patrimoine (Heritage Code). It then takes into account the technical limits arising from web archiving methods. Finally, after having determined what has to be collected, and then what can be collected, it applies an economic approach to identify what in particular it seeks to collect. The BnF has therefore been led to define principles of selection within the context of legal deposit; to this end, each department involved in web archiving has developed, by means of much uncertainty and many experiments, its own collection policy. The BnF must now undertake the task of combining these different policies as part of the revision of its collection development charter.

1) Defining the scope of the collection

The legal restrictions

The scope of legal deposit is of course derived from the law. The legal framework was established in two phases: the law which extended legal deposit to the internet was voted into law in 2006; the enabling decree was published in 2011 [1]. The law defines the object of the legal deposit in very general terms: included are all "signs, signals, writings, sounds or messages of all kinds which are the object of diffusion to the public by electronic means". In the spirit of legal deposit, all online publications are thus included in the scope, be they written text, videos, online games… provided that the content does not have the status of private correspondence. Private sections of social networks are thus excluded from the collection, while public pages are included.

The decree adds a certain number of specifications[1]. Firstly it defines what should be understood by the "French internet": starting with all sites hosted on French top-level domains (TLDs), .fr, .paris, .re for Réunion Island, etc.; and/or sites whose domain name has been registered by a person resident in France; and/or sites produced on French territory.

[1] http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006074236&idSectionTA=LEGISCTA000025005272.

In addition, the decree defines the distribution of tasks between the two depositary institutions. The National Audiovisual Institute (Institut national de l'audiovisuel, or INA) receives the responsibility for television and radio websites, as well as all sites that are "principally dedicated" to these media; the scope of the BnF is for its part defined by default: all of the French web… except the sites belonging to the scope of INA.

Finally, the decree defines the practical aspects of the means of collection: all domain names must be collected, however the depth of collection is not specified, and there is no requirement to perform an exhaustive collection of each site. The decree also defines a minimum archiving frequency: once a year – leaving the depositary institutions the possibility to collect certain sites more often.

The technical restrictions

In order to fulfil its mission, the BnF – like many other institutions across the world – uses "robot" technology: this is in fact software that, starting from a given list of URLs, roams the web from link to link discovering and collecting the content that it has been asked to harvest[2]. These robots make it possible to archive large quantities of data but they have their limits. They are poor at capturing the "deep web": databases and highly interactive documents are beyond their capabilities. Videos too are often poorly collected, unless specialised and intensive archiving processes are implemented. These limits form a second restriction on the scope of the collection: the subject of legal deposit is, first of all, everything that the depositary institution is capable of capturing.

An economic approach: the mixed model

Even after taking into account these two kinds of restrictions, the field to be covered remains extremely vast: the size of the French web is estimated at seven or eight million sites, some of which have a very high rate of change[3]. Comprehensiveness is therefore no longer an achievable objective, if only for reasons of cost and economic sustainability. This ideal has therefore been replaced by the objective of representativeness: the aim is to form an image that, while incomplete, is faithful to the nature of the French internet, taking into account all types of publication and all kinds of content, from the most serious to the most insignificant.

To this end, the BnF combines two models of harvesting. The first is the "broad crawl", performed once a year. This crawl covers all the sites that the BnF has been able to identify as being French, currently more than four million. The depth of the crawl is limited (several thousand files per domain name), but is sufficient to collect in a satisfactory manner more than 90% of sites identified. This harvest is complemented by "focused crawls" which concern sites that are to be collected more often (up to once à day) or to a greater depth (up to several hundred thousand files per domain); also included are sites to be collected in relation to a given event (elections, festivals, sporting occasions…). The identification of these sites is performed in cooperation within the BnF by a network of "corresponding officers" who are responsible for the selection, the quality control and the promotion of the sites that they have asked for. These corresponding officers are divided between a dozen departments covering thematic collections (Literature and Art, Science…) or legal deposit (Maps, Prints, Audiovisual material...). In the case of specific projects, these personnel are sometimes joined

[2] The majority of the tools used for crawling and access to the web archives at the BnF are open source and have been developed in the context of the International Internet Preservation Consortium (IIPC). See http://www.netpreserve.org/.

[3] AFNIC estimates that the .fr TLD represents more than a third of French sites; there are currently 2.7 million sites on .fr (http://www.afnic.fr/data/actu/public/2010/afnic-french-domain-name-report-2010.pdf, p. 7).

by librarians or researchers in partner institutions (regional or university libraries, research laboratories, etc.).

Finally, in accordance with its objective of economic sustainability, the BnF defines each year the maximum size of the collection that it has the capacity to build and conserve in the long term; in 2014 the annual volume was thus fixed at 100 terabytes (TB). Out of this total, 50% of the resources are assigned to the broad crawl and the other 50% to the focused crawls. The BnF selectors must therefore take account of a "budget" which is defined in terms of files or bytes rather than euros.

To aid the selectors in their work, a selection tool has been put in place called "BCWeb" or "Building Collections on the Web". For each site, it is possible to indicate the URL, keywords and crawl settings. The latter are of three kinds: the frequency of harvesting (daily to annual), the depth (the whole site, part of a site, a single page…) and the "budget", i.e. the target size in terms of the number of files.

These settings are essential as they allow a precise definition of the object of the crawl. They are themselves the result of an economic approach: it is necessary to define, for each site, whether it should be collected very often, even if it means settling for only the pages directly linked to the home page (as is the case for news websites), or rather in greater depth but infrequently.

This economic approach thus aims at defining production models for websites that are standard and shared: official or academic sites, news, blogs… However, these models do not say anything regarding the content of sites, the choice of which remains at the discretion of the different departments involved in selection. It is at their level that, during a period of progressive maturing, collection policies have been defined.

2) Collection policies for focused crawls

Founding principles

The focused crawls are aimed at overcoming the principal shortcomings of the broad crawl. It is therefore this kind of crawl to which the concept of collection policy best applies, with all that it implies in terms of selection and intellectual construction used to complement the mass legal deposit constituted by the means of the broad crawls.

The documentary strategies for web legal deposit that have been progressively put in place by the institution are in accordance with certain major founding principles of the BnF, affirmed in its collection development charter in 2005:

- France is a priority object of study: while the broad crawl covers domains registered in France, the focused crawls are also strongly concentrated on French sites. Nevertheless, this territorial dimension can seem problematic, or even diametrically opposed to the web which is by nature hardly divided up by national boundaries, and some librarians have undertaken, in a very limited way, the collection of foreign sites that are complementary to the French domain, following the model of the acquisition of foreign publications on physical media.
- Encyclopaedism: all domains of knowledge are represented, as is the case with all collections at the BnF, be it language and literature, social science and humanities, science and engineering, arts…

- The temporal dimension: the objective is to combine harvests “in the long term”, which aim as complementing existing collections on all media, with crawls closely linked to current events. Thus, in 2013 crawls were specifically dedicated to the war in Mali, the law on marriage for same-sex couples and the election of the new Pope.

Ongoing crawls

Since the internalisation of the harvests (2006), the network of corresponding officers for web legal deposit has developed collection strategies by adapting approaches that are complementary rather than contradictory: selection/sampling, continuity of collections/exploring new territories.

Two approaches can be identified,

> *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. [1]

It is thus possible to observe schematically two kinds of reasoning: that of departments that acquire uniquely by purchase, where a culture of selection reigns that is intended to develop a collection aiming at excellence not only in each of its elements but in its structure; and that of the departments that handle legal deposit, which have experience in collections aiming for exhaustiveness[4].

In addition, the web is approached from two different angles: the continuity of collections and the exploration of new territories.

The ongoing crawls aim to bring together, in an approach that prolongs that of the acquisition of physical material, a literature that is increasingly varied and voluminous, and now largely in digital form: official publications, sheet music, maps, music, cinema, periodicals and ephemera of all kinds. The internet is also a domain where publication is more flexible and less expensive than that of physical media, thus making possible the creation of all kinds of new material which represent potential sources for research: grey literature, disciplines that are under-represented in printed publications….

[4] At the Bibliothèque nationale de France, the collection and treatment of documents received by legal deposit are organised by the type of document and handled by five departments: the Legal Deposit Department for printed material (books, brochures, periodicals) and online digital material ; the Audiovisual Department for sound, audiovisual and multimedia material as well as digital material on physical media; the Prints and Photography Department for prints, photographs, posters and images; the Music Department for printed music (sheet music); the Maps Department for cartographic documents (monographs, maps, plans, globes, atlases, etc.)

But the web is also considered for itself, as a source and an object of study in perpetual evolution, with its own characteristics: hypertextuality, graphical innovation, interactivity, real-time updates…

By examining more closely the focused crawls put in place by the different departments of the BnF, it is possible to observe that the strategies put in place combine these different approaches, in order to "cover" the range of possibilities offered by the web.

A first example is the harvest of news websites. On the one hand, the logic of continuity of collections is fully represented in the will of the BnF to collect, by means of robots, the PDF versions of local editions of regional daily newspapers[5]. On the other hand, special attention is given to the new forms of online news: pure players such as Médiapart and Rue89, portals that pass on information from news sites…

Another area where this tension between continuity and novelty can be seen is that of maps: more than 300 sites are the object of focused crawls in the domain of cartography and itineraries, from the most well-known institutional sites to amateur blogs. For the BnF Maps Department, it is a question of maintaining the continuity of its missions of preservation of physical material, to conserve a trace, which is not fully exhaustive but rather representative, of the manner in which space is represented at a given moment, and the uses which can be reproduced in a cartographic manner [3].

In the same way, in the domain of the performing arts, the exhaustive approach is combined with that of selection: the websites, generally small in size, are relatively well collected by the broad crawl, allowing an important sampling of what is produced on the French web. The focused crawls thus have as their objective the enrichment of these archives. The selection of websites has been created by successive layers: starting with the institutional sites of theatres that were poorly documented, in order to complement the documentary holdings by means of the web, followed by the other specialities of the department. A particular attention was given to theatrical criticism, thanks to a selection of blogs. Information websites unconnected to any theatrical structure, sites that make sources available online, and in general sites that were not the transposition or the continuation of what was being produced on another media, were later added to the selection. Finally, in 2013, a new group was added, a dozen sites related to archive collections held by the Performing Arts Department. Performing Arts currently represent more than 180 selected sites [4].

Finally, one last example: the ongoing crawls concerning audiovisual material were conceived around three main approaches. First is the the continuity of the collection in relation to audiovisual and multimedia documents, to archive the sound, audiovisual and multimedia material distributed on the internet, the producers and publishers of which were traditionally (or still are, in parallel) depositors for the legal deposit of phonograms, video recordings or multimedia documents on physical media. By extension, this principle of continuity is applied to websites that distribute sound, audiovisual or multimedia material which would previously have been distributed on physical media, but where the producers or publishers have started directly with the internet. The second approach used consists in collecting those sites that represent innovative forms of creation and/or distribution of sound, audiovisual or multimedia material which have appeared with the internet, such as sites for

[5] For cost reasons, the paper version of local editions of regional daily newspapers are no longer kept by the BnF. For more information, see [2].

sharing content produced by users, Net Art websites… Finally, the crawl collects, as an extension of acquisition of physical media, sites that document the universes of music and sonic creation, cinema and audiovisual material, video games and digital creation. Audiovisual material, combining the desire for exhaustiveness, the reference approach and the representative approach, currently represents almost 5 000 selected sites [5].

Project crawls

In addition to ongoing crawls, there are also project crawls. These have a more narrow scope than the ongoing crawls, both in terms of subject and in time. They answer a defined documentary need, with a limited scope, and by nature are not intended to become permanent. They are in general the product of cooperation, not only between departments within the BnF, but also with other partners.

The first project crawl, put in place in experimental form as early as 2002 and made systematic after the law of 2006, concerned electoral campaigns [6]. These campaigns are in effect played out as much on the web as in the streets. The websites of political figures and activists constitute a valuable resource for understanding the issues and the results of an election, but are by nature extremely volatile. The BnF has therefore put in place a system of selection, archiving and access to this new kind of source, and has applied it to most of the national and local elections held since 2002. These crawls have in particular allowed the documentation of the use of new technologies by political parties: thus, political blogs dominate the campaign in 2007; the use of social networks such as Facebook and Twitter is timid in 2009 then becomes more widespread after 2010.

Another example of a project crawl is that covering diaries which, with the fashion for creating blogs, have undergone a considerable revival. To conserve “the huge and unprecedented field of autobiographical expression that has been created”, web archiving has become a necessity, with selection performed by the Literature and Art Department of the BnF for blogs and websites of writers, and by a partner, the Association for Autobiography and Autobiographical Heritage, for websites of personal expression [7].

From selection to giving access

The articulation of different kinds of harvest, combining large-scale crawls and more focused human selections, makes it possible to answer the expectations expressed by the public during a prospective study on the representations and the expectations of potential users of the web archives, performed by the BnF in 2010-2011. While professionals and the “average user” of the Research Library expressed only a relative and generally isolated need to use web archives, researchers, who recognised both the great richness and the volatility of the web universe, expressed the difficulty of defining and circumscribing significant corpora. Faced with this difficulty, the researchers clearly perceived the BnF as a trusted third party capable of guaranteeing an access for researchers to collections that are reasoned and documented. In terms of content, the expectations of researchers indicated several possible approaches: developing crawls to preserve the traces of important nodes and networks; archiving the most popular sites and also those that break new ground; finally, to collect not only isolated, discrete elements but keep trace of the practices on the web that mark the spirit of the times and document social, commercial or other trends on a large scale. In any case, for all of the categories of users questioned, it was clear that the concept of selection by librarians was perceived as legitimate, inevitable and necessary given the volume of data to be archived [8].

The expectations of the public also included a more practical way of accessing the collections thus constituted; while access is restricted by the law to the interior of the “research” levels of the BnF, progress can be made concerning the offer of effective access tools that take into account the hypertextual dimension of the internet. The Wayback Machine, used in the main consultation interface at the BnF, allows researchers to navigate within the archives as they would have done on the live web. This spatial navigation is enhanced by a temporal exploration. Starting from a given site, it is possible to go back in time and analyse its successive transformations.

This kind of indexing presupposes that, to discover a site, its address is already known… To mitigate this problem, it seems indispensible to develop full-text indexing which allows users to search for pages and files depending on their textual content using keywords. Due to the huge volume of the collections (21 billion files, 470 TB of data), and the difficulty of handling multiple temporal layers, this full-text indexing has so far only been performed in a very limited fashion, which represents a real obstacle to the use of the collections.

In parallel, in order to overcome the difficulty for the uninitiated of distinguishing what is proposed in the web archives from that directly accessible in the web, the BnF has put in place a presentation in the form of “Guided Tours”, conceived as flagship products on subjects which can be easily understood and which are representative of national, political and cultural memory, where the phenomenon of disappearance is clearly apparent.

3) What next?

Integration in the collection development charter

The BnF has undertaken a revision of its collection development charter, which dates from 2005. For the revision of the charter, which is due to be completed by the end of 2014, several methodological choices have been taken: to approach collection policy as a whole, in terms of collection enrichment whatever the mode of entry or the media; to prioritise the presentation of large disciplinary fields, more understandable to the public, rather than by the organisation in departments. The revised charter will include an introductory section devoted to a reminder of the context and a synthesis of the major evolutions over the last 20 years, and perspectives until 2020. The main part of the charter will consist of thematic sections intended to present, by disciplinary field, the strategic priorities for collection development at the BnF.

The collection development charter for acquisition by the BnF, published in 2005, did not take into account web legal deposit. Its revision, currently underway, now includes it: because of the position it has taken among the activities of BnF staff, because of the volume of data collected, because of its functioning costs, and especially because of its inescapable role in ensuring the completeness of heritage collections, and therefore of the missions of the institution as they are described in its founding decree. Internet legal deposit will therefore be the subject of a particular focus in the introductory section, in order to define the problems its poses, and will then be included in each thematic section as the collection policy of the BnF is now imagined as a global strategy of collection development, whether on physical or digital material.

Towards an affirmation of content priorities

Web legal deposit, on the scale of the lifetime of the BnF, is part of its very recent history. It has been put in place following a solid and reasoned outline that is both technical (the “mixed model") and organisational (the network of corresponding officers), and builds on structural and shared principles and tools.

For the different actors in the library, this has required a period of training and adaptation, to approaches, typologies and technical constraints to which the librarians were not accustomed, but it is now accepted that the different decision taken by a web legal deposit corresponding officer during the process of selection are indeed part of the skills of a librarian, in the sense that they guarantee the quality, the representativeness and the relevance of the resources collected.

Nevertheless, several observations must be made: apart from the fact that the human resources put in place by different departments are very unequal, the ideas of costs, as with physical documents, are very different depending on the domain: in a simple sense, a video "costs" more (in bytes) than a site that is largely textual; finally, the crawls put in place by different partners do not have the same "weight" in terms of number of bytes depending on whether the logic is principally that of exhaustiveness or selection.

Since the internalisation of the focused crawls, there has been a deliberate choice to not set a defined volume by subject or by department. However web archiving must function within a necessary control of costs, the risk being that of collecting ever more sites but ever more superficially, of favouring quantity at the cost of the depth and thus the quality of the collection.

Collection policy traditionally includes in its reflection the idea of elimination of collections. However, in the current case, that which is collected under web legal deposit is by definition destined to be preserved, to "become heritage". But the network of selectors has a mission of constant surveillance in order to adapt the crawls as precisely as possible: reduce the frequency of sites that become less active, stop the crawl of a site that no longer seems relevant, add newly-created sites to the crawl… It is not therefore a question of discarding collected material, but an implicit exclusion beforehand, that applies to web archiving.

While the annual volume (in TB) allocated had up till then been sufficient to perform all of the focused crawls as wished by the corresponding officers, 2013 was the first year in which it became necessary to make exclusive choices in order to ensure that the crawl was as diverse and as representative as possible. It was therefore decided to stop project crawls that were considered complete, to limit the frequency of focused crawls where the greater or lesser vitality of sites allowed, to perform the task of removing sites selected in duplicate in crawls led by different actors, and to limit the crawl of videos to 10TB per year.

After a period of technical, organisational and intellectual maturing, it now seems necessary to collectively affirm the documentary priorities: as is often the case, the preoccupation with collection policy comes in the first instance from budget constraints (in this case, the archiving capacity in TB), and makes it necessary to ask questions in terms of long term strategy, asking which collections have the strongest added value for future research.

That said, how does one identify today what will interest researchers in the future? The recent opening conference of the IIPC, held in Paris in May 2014, demonstrated this with the presentations of several researchers: the fields of research are numerous, each project using different models of corpus constitution and analysis tools, and the inventiveness of researchers is as large as the richness of the web. It is therefore vital to "manage the future" by maintaining the combination of different approaches to collecting the internet in the service of a measured encyclopaedism.

References

[1] Illien Gildas, Sanz Pascal, Sepetjan Sophie, Stirling Peter, “The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future”. In : *IFLA journal*, 2012, vol. 38, n°1. [http://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-38-1_2012.pdf, consulted 20th March 2015]

[2] Oury Clément, “When press is not printed*:* the challenge of collecting digital newspapers at the Bibliothèque nationale de France”, *Preconference of the IFLA newspapers section, Aug 2012, Mikkeli, Finland.* [http://halshs.archives-ouvertes.fr/docs/00/76/90/84/PDF/LegalDepositNewspapersBnF_Oury_IFLA2012.pdf, consulted 20th March 2015].

[3] Lebailly Guillaume, Marchand Emmanuelle, « Cartes et carnets de voyage en ligne : la BnF collecte le Web », *La géographie*, 2013, n°1549, p. 42-43.

[4] Obligi Cécile, « Les archives du web à la BnF : un formidable gisement à venir exploiter ! », *Skén&graphie*, n°2, Annales littéraires/Presses Universitaires de Franche-Comté, Besançon, 2014.

[5] Carou Alain, « Archiver la vidéo sur le web », *Bulletin des bibliothèques de France*, n° 2, 2007. [http://bbf.enssib.fr/consulter/bbf-2007-02-0056-012, consulted 20th March 2015].

[6] Oury Clément, « Soixante millions de fichiers pour un scrutin. Les collections de sites politiques à la BnF », *Revue de la BnF,* 2012/1 n° 40, p. 84-90. [http://www.cairn.info/resume.php?ID_ARTICLE=RBNF_040_0084, consulted 20th March 2015].

[7] Genin Christine, « Collecter l’océan ? L’archivage de l’intime en ligne à la BnF », *Bibliothèque(s)*, n°47-48, décembre 2009, p. 50-52.

[8] Chevallier Philippe, Illien Gildas, Stirling Peter, « Web Archives for Researchers: Representations, Expectations and Potential Uses », *D-Lib Magazine*, 2012, vol. 18, no 3/4. [http://www.dlib.org/dlib/march12/stirling/03stirling.html, consulted 20th March 2015].