Web Archives for Researchers: Representations, Expectations and Potential Uses

Peter Stirling, Philippe Chevallier, Gildas Illien

Web Archives for Researchers: Representations, Expectations and Potential Uses

Peter Stirling, Philippe Chevallier and Gildas Illien
Bibliothèque nationale de France, Paris
{peter.stirling, philippe.chevallier, gildas.illien}@bnf.fr

doi:10.1045/march2012-stirling

Abstract

The Internet has been covered by legal deposit legislation in France since 2006, making web archiving one of the missions of the Bibliothèque nationale de France (BnF). Access to the web archives has been provided in the library on an experimental basis since 2008. In the context of increasing interest in many countries in web archiving and how it may best serve the needs of researchers, especially in the expanding field of Internet studies for social sciences, a qualitative study was performed, based on interviews with potential users of the web archives held at the BnF, and particularly researchers working in various areas related to the Internet. The study aimed to explore their needs in terms of both content and services, and also to analyse different ways of representing the archives, in order to identify ways of increasing their use. While the interest of maintaining the "memory" of the web is obvious to the researchers, they are faced with the difficulty of defining, in what is a seemingly limitless space, meaningful collections of documents. Cultural heritage institutions such as national libraries are perceived as trusted third parties capable of creating rationally-constructed and well-documented collections, but such archives raise certain ethical and methodological questions.

1. Introduction

Under the law passed on the 1st August 2006, one of the missions of the Bibliothèque nationale de France (BnF) is the legal deposit of the French Internet.[1] The public consultation of the web archives, which remains experimental, was first introduced in April 2008; starting with a handful of terminals, it has been progressively extended to all the computers in the reading rooms of the research library of the BnF.[2]

The use of the web archives remains limited. Access statistics show that there are 30 to 50 connections per month; this level of use has remained stable for the past two years, but represents only a very small fraction of the total number of people who use the library. A small number of research projects have already used the BnF web archives;[3] in these cases the researchers have often made contact directly with the Digital Legal Deposit service at the BnF.

In late 2010/early 2011 a study was carried out by the Strategy and Research Delegation of the BnF, whereby potential users of the BnF web archives were interviewed with the aim of exploring their needs both in terms of content (the collections available in the archives) and services. In the original plan it was hoped to compare actual users with those who had not yet consulted the archives. In fact this proved impossible, largely due to the limited number of actual users: a message was included on the consultation interface to invite people to participate in the study but this did not produce any responses. At the same time, the methodological choice was made not to

contact users who were already known to the Digital Legal Deposit service, to avoid any bias. The final study is therefore focused uniquely on potential users rather than people who have experience of using the archives. This approach allowed us to explore ways of increasing the use of the archives by understanding the needs and expectations of people who do not currently use them.

This was a qualitative study, which was intended to analyse both the subjects' practice in web research and their ways of representing the web archives. Fifteen interviews were held with three user groups: researchers, professionals and "average users" of the research library at the main location of the BnF in Paris, with the interviews being recorded and transcribed for analysis.[4] In this article we concentrate on the first user group, consisting of five researchers working in various areas relating to the Internet (history, philosophy, sociology, information technology). Of these five researchers, three were aware of the existence of the BnF web archives but had not yet consulted them.

As the subjects were removed from any actual, or even potential, use of the archives, the interviews were necessarily general in tone, which gave rise to fundamental questions regarding the collections and the definition of the web archives. Rather than asking questions on the services currently provided by the web archives of the BnF [5] (such as the kinds of collection, their frequency, access tools, etc.), it seemed more interesting to examine in depth the subjects' own practices in terms of research and archiving in relation to web resources, to provide material that will allow the BnF to reflect on the collection policy and the tools and services that should be put in place for users.

While the interest of maintaining the "memory" of the web is obvious to the researchers, they are faced with the difficulty of defining, in what is a seemingly limitless space, meaningful collections of documents. Cultural heritage institutions such as national libraries are perceived as trusted third parties capable of creating rationally-constructed and well-documented collections, but such archives raise certain ethical and methodological questions. The article aims to provide a synthesis of the different ideas raised by the interviewees. It starts by examining the use of the web in research, before moving on to web archiving and the perceptions of it held by researchers. It then examines approaches to constituting meaningful web archives, and data and link mining along with mapping tools and services that could be offered to respond to researchers' needs.

2. Use of the web by researchers

The web is today an everyday reality for social science researchers. It exists not only as a means of accessing scientific literature but also as a field of research, more or less formal, in itself; in addition, researchers themselves also have an online presence, on social networks and blogs. Despite this recent opening up of the web as a field of research, the use of online documents in scientific work remains problematic, due to the difficulty of creating "corpuses" that are sufficiently controlled, organised and shared by the research community.

Taking the temperature of the web

Independently of research projects that are specifically focused on the web,[6] the web is frequently cited as a "starting point" for research work. It provides a first impression, a walk-through of a subject, which often lacks any defined methodology and is linked more to the personal web habits of the researcher. This period of exploration is particularly used by researchers exploring new

subjects, and especially in areas that are innovative and shifting (urbanism, design) or those that create public debate and controversy (paedophilia, public order).

The words used by the interviewees to describe this activity emphasize the "newness", but also the lack of importance of the material encountered: "the latest thing", "little things", "little papers". These "little papers", as opposed to full academic papers, are often to be found on blogs: those of researchers, but also consultants, amateurs or actors in a given area. A heterogeneous collection of sources is thus created, where objective information mixes with opinion and speculation. This involves what one interviewee calls "a lot of blabla" — but this "blabla" is exactly what allows researchers to study how their subjects of research are talked about outside the academic community, and to examine the difference between the academic discourse on a subject and the way it is talked about in wider society.

Monitoring and filtering the web

To find source material on the web, while some researchers continue to use Google (although recognising its limits), others have developed filtering strategies: rather than searching the web directly, they set up a "surveillance" of the web by selecting sources to follow. This "filtering" process today also uses social networks such as Facebook and Twitter, which are replacing RSS feeds: according to one interviewee, it is necessary to sort through RSS feeds, whereas Twitter is seen as being more direct, as bloggers signal immediately the latest posts on their site.

However the brevity of tweets gives rise to multiple uses, and it can be difficult to distinguish between a researcher's professional and personal use: "It's difficult to say at a given moment whether I'm performing scientific monitoring or just tweeting with my friends". The action of monitoring the content thus leads to the web becoming a place where the researcher is used to showing him- or herself, in a series of widening concentric circles: Facebook profile, Twitter account, blog. The web provides researchers with another way of presenting themselves, by sharing the results of their monitoring (even while hiding their sources, so as not to lose the advantage) or by promoting their non-academic writing via blogs. New codes of recognition are developed, for various reasons. Firstly, the researcher-blogger still has a "bootlegger" aspect and operates frequently "in secret", hidden from certain colleagues who think that such writing is not "serious". In addition, blogs allow researchers to experiment with new ways of writing that may be far removed from their published papers.

Problems of using the web in scientific work

The use of the web which is principally examined in this study is not that of citing an article published online, which has become usual practice for researchers, but rather the use of a website as evidence or illustration of a sociological or historical phenomenon. Several of the researchers questioned recognise the difficulty of using this kind of material in their work, due to the lack of an established methodology. This comes down to two basic problems. Firstly, there is the need to be able to cite a source which can be re-examined independently by other researchers, particularly problematic in the case of sites that may change their address, or disappear completely.[7] Secondly, there is the difficulty of being able to justify the selection of a given site to be used in the research: there is a need that a source should be part of a defined corpus that is shared by a community of researchers. This means that any use of web sites must be documented and justified, to show why one site has been used rather than another.

However, no existing expertise about the web seems to allow such a justification; someone who really knows the web knows the limits of their own knowledge. This comes largely from the fact

that sites are organised by networks and each researcher creates their own web based on the sites that reference each other — meaning another important site in the same area could remain completely unknown if the researcher is not in the right "circle". In these conditions, the contours of the corpus to be studied remain uncertain. In the area of net art, for example, it is often necessary to search for artists' names in search engines, and discoveries can often only be made by accident, as it is often a closed circle. Since it is impossible to be exhaustive, even in a relative manner, any analysis of novelty or anteriority becomes relative, as it is difficult to say definitively if a site is really "new", or if another site showed the same features three hours or three years before.

The documentary basis is not seen as solid enough to allow a verifiable academic work, as the knowledge on which it is based is insufficiently shared. The web is even seen by some teachers as not to be recommended to PhD students, as they could miss something important. For research work on the live web, the researchers have however managed to put in place some protocols to define corpus that will be as representative as possible:

- Define the perimeter: for example a cultural boundary such as a nation, or a language ("the French-speaking web");
- Explore and observe: even if the observation uses search engines, determine the most efficient search terms to define a first nucleus of sites;
- Selection and qualitative description: this is based largely on links between sites, but the links must be qualified depending on their type; this is therefore manual work. This allows for example a cartography of the blogosphere and its structure, and a comparison of different blogospheres (in different countries, languages...)

Despite this kind of protocol, in the case of volatile communities that may pass quickly from one technology or platform to another, the data collected on the web are still seen as unreliable and ephemeral, and the researchers themselves recognise that it is necessary to collect frequently sites that interest them.

Data collected on the web

Research projects that relate directly to the web and its usage therefore require researchers to not only search for but also store data existing on the Internet. One interviewee distinguishes three types of data, noting at the same time that they are not necessarily to be conserved indefinitely:

- Qualitative data: online content to be analysed for its meaning, "from a lexicographical point of view". This could be images or texts.
- Quantitative data: figures and calculations, such as connection time, number of contacts, number of readers for a given site.
- Relational data: a group which is beginning to have a considerable importance but which creates lots of problems: "[it] also makes us face questions that didn't exist before: who talks to whom? Who is connected to whom?" To do this, it is necessary to "cluster, analyse the networks, see what are the degrees of density or the kind of interconnections that can be created among friends' networks". But such an analysis also poses ethical problems: "because it's really difficult to make relational data anonymous".

Independently or on the margins of research programmes where the storage of such data is performed by an institution or company,[8] some researchers have attempted to create their own personal archives of websites. This often involves the use of printouts or screenshots; other researchers, particularly those working exclusively with the web, create their own archives of files which can reach several gigabytes and on which they use their own textual analysis tools. In this

second case, a web archive is only considered useful in providing data "to make the machine work". This approach is however seen as very costly both in terms of time and equipment. Self-archiving, whether by printouts, screenshots or downloading, also creates practical difficulties: interviewees speak of offices overwhelmed by paper or hard drives.

Such practices respond to a sense of urgency faced with the volatility of the web. However this kind of archiving, using printouts or screenshots, is a stopgap measure, and researchers who use it recognise its lack of established principles or scientific validity. One says: "The problem is, these pages are updated. So the fact that you've printed out these pages, from the point of view of the authenticity of the source, means nothing at all. Printouts or screenshots only give static material, whereas the sites are meant to be updated and have dynamic content".

For researchers who do not have a systematic practice of web archiving, the use of Internet Archive[9] is common. Such usage, although it may be intense during certain periods, remains occasional, and is often caused by the disappearance of a site that had been used in previous research. One researcher admits having abandoned using Internet Archive several years ago, due to the absence in many cases of dynamic content, and many broken links to other sites. The researchers in the study who were aware of the harvesting of French sites by the BnF (either from the BnF itself or by a message from their institution) had never visited the library to consult the archives. They do not feel an immediate need for these archives, but can imagine that they will become useful for their students, or for projects more directly linked to the history of the web.

3. Perceptions of web archives by researchers

Before entering into the details of the recommendations of the researchers in terms of what should be included in web archives, it is important to examine their general perceptions of web archiving: the interest, the legitimacy and the manner of approaching such an activity.[10] It also becomes clear that it is necessary to find a way of adequately representing this new kind of archive: to what can it be compared?

The value of archiving the web

All the researchers interviewed recognise the value of maintaining the memory of the web; as we have seen, they recognise the volatility of the Internet and most have themselves experienced problems in their research due to the disappearance of a website, leading to the archiving practices described in the previous section. A notable example is that of online art, where the earliest examples, from the very beginnings of the Internet, are now known only from descriptions in books as the originals are irremediably lost. The researcher who cites this example notes that, to his knowledge, no-one has archived early examples of forums and newsgroups where early Internet art, such as ASCII-art,[11] circulated.

This volatility can be particularly marked in certain areas of the web cited by researchers: sites that are judged undesirable by ISPs (pro-anorexia...), communities created by young users who move rapidly to the newest site or technology (WebRing to Tumblr...) or sites that are linked to underground or illegal activities (Undernet...). Blogs are perceived as being particularly fragile, as they are maintained by amateurs who may not have the time or the means available to devote themselves to their passion for an indefinite period.

The disappearance of online material may also be intentional on the part of its authors, in particular in the case of institutions that wish to remove the evidence of developments in their views, to give the impression of an unchanging position. This can give a "political force" to a web archive, as shown by the example one researcher cites of a French national agency that changed the vocabulary of its website overnight without any communication: the researcher discovered this using Internet Archive. Such a web archive therefore allows researchers to put into perspective the latest version of a website, which their students often take for the definitive version, and allows the distance necessary for a historical approach.

While the interest of web archiving is thus acknowledged, there remains a confusion arising from differing representations of what an "archive" of the Internet may be. Due to the existence of an "archives" section on many websites, or the history of changes on a site such as Wikipedia, the web often gives the impression that it functions as its own archive. From the point of view of the historian, however, a true web archive has to maintain the trace of a previous state of the site, where the contents are presented in their original context: the architecture and layout of a site is part of its means of expression.

Ethical questions arising from web archiving

The legitimacy of web archiving is viewed differently by the researchers depending on the kind of site to be archived, leading rapidly to questions about the right to privacy and the definition of what is "public".

Archiving is most clearly seen as justified in the case of government and institutional websites. These sites have an apparent stability and often include an element of online publication of reports and other documents; indeed the vocabulary used by the interviewees shows that these sites are readily considered as "publications" themselves, as opposed to other sites that are far from having this status unequivocally. This category of sites is thus most often cited in propositions of what to collect, for two reasons. These sites seem to resist the unstable nature of the web, are easier to circumscribe and therefore to capture. At the same time, it is this apparent stability that can be analysed using web archiving: as the example above shows, web archiving can show how official positions can be contradicted and changed without warning. The right to privacy is not seen to apply to the public debate, and web archiving can allow the vigilance necessary in a democracy.

Blogs are also considered suitable for archiving. The researchers consider that they are written to be read, and represent a particular form of writing that deserves to be preserved; in addition, most blogs already maintain their own archives. This approach was strongly related in the study to the practice of the researchers interviewed, who were in favour of their own blogs being preserved and presented in this way.

Archiving material such as that on Twitter or social networks such as Facebook, on the other hand, gives rise to more questions. The researchers consider that this leaves the domain of publication and becomes more that of conversation, which is intended for a limited circle of people; the definition of how "public" this circle may be is difficult. In addition, such social networks allow a degree of anonymity, or at least, by the use of multiple pseudonyms, people can behave and communicate differently with different circles — friends, family, colleagues — just as they would in "real life".

This concept of the Internet as an area where people interact creates a problem regarding its archiving. The Internet is no longer just a place for publishing things, and large parts of the web, even those publicly accessible, may be considered not as published "information" worthy of

archiving, but rather the traces left by actions that people could equally perform in the streets or in a shop: talking to people, walking, buying things... It can seem improper to some to archive anything relating to this kind of individual activity. On the other hand, one of the researchers acknowledges that archiving this material would provide a rich source for research in the future, and thus compares archiving it to archaeology, requiring analysis similar to that of studying different layers in the earth. At the same time, the same researcher says that such a "Borgesian" project is "completely terrifying".

As well as doubts about the collection of such material, there are also questions about possible improper use of the data. One researcher stresses that the BnF must use a system of accreditation, and in particular control the use that is made of data, on the model of that used for databases in laboratories where all exports of data can be traced.[12]

Ways of imagining the archives and of defining (or not) their constituent parts

The very idea of an archive is compared to, and opposed to, the idea of the "flow" of the web: "how do you archive the flow of time?" asks one researcher. The web works as a continuous flow of material, and this aspect has tended to increase in recent years: whereas previously sites consisted largely of static content updated at intervals, today the use of streaming and interactivity means much more dynamic content: researchers cited examples such as dating websites and Chatroulette.

The traditional paper archive is mentioned more than once as a counter-example to describe the web archives, which cannot be considered like a "19th century" archive with stable fixed content: the archive has to be as dynamic as the material it seeks to preserve. By definition, a flow does not contain individual parts or units, and it requires an external act to separate it into "sequences". In seeking to define a structure for the web archives, even the website is not necessarily relevant as a unit, as websites are defined by their place in the network.

Rather than trying to apply old models of archiving, one researcher suggests that it is necessary to find new models to describe this new kind of archive, and suggests that it should be thought about in terms of the study of oral traditions rather than the printed record, and that archiving practices imagined with this in mind could better allow the dynamic and temporal nature of the web to be captured.

4. Content and selection in web archives

Collection policy: don't be scared of selectivity

Despite the importance of preserving the links between websites, it is necessary to "cut" these links to allow the preservation of what is otherwise an infinite space. The lack of selection in what is archived is seen as unthinkable when faced with the web. One singularity of the web is that it seems to abolish most of the selection procedures that are applied in print-based publishing, and which therefore affect print legal deposit. This can create the illusion of an undifferentiated surface, when in fact the content published online is not all of the same value or for the same audience. The web forces us to reconsider the question of what should and should not be preserved.

This kind of selection can in fact be compared to traditional archiving techniques, where all preservation is predicated on selection of material to preserve and destruction of the rest. The researchers, however, are divided between an acknowledgement that everything could be

interesting for future historians, and the awareness that choices must be made in creating a meaningful archive. The researchers accept the idea of not having everything, on the condition that the rationale for the choices made is clearly explained. Historians are used to dealing with incomplete archives. However it is necessary to describe and justify the criteria used, as discussed below.

One researcher had reservations about automatic collection as it presented the possibility of automatic treatment of the data, which may include personal data. Indeed only one of the researchers interviewed stressed the importance of making a large, "relatively random" collection at a given moment, to provide a panorama of what exists.

Possible approaches: by nodes, by theme, by "novelty", by practice

The researchers interviewed did not know in detail the current practices used to collect Internet material at the BnF,[13] which allowed them to propose various approaches:

- Collecting starting from pre-determined "nodes", based on a search that could use mathematical models. This should allow a "selection" based on the audience, to collect the sites that are most visited or most linked, to give an idea of what was important at a given time.
- Sampling, based on themes, by kind of site, or kind of activity.
- Focused collections should concentrate on what is really new and groundbreaking, possibly linked to research programmes that will identify sites that "renew a genre" on the Internet. However a small sample of a given site could be enough (for example Chatroulette was cited as being "repetitive").
- Another researcher proposed the idea that the Internet is a "practice", based primarily on the use of search engines such as Google. To reflect this, collections could be based on the Google results for terms based on a research project, or chosen at random in a dictionary.

Finally, the question arose about the relationship between selection policy and legal deposit. The term "Internet archives" seems incongruous in the context of legal deposit since, as mentioned above, the practice of archiving is based on selection and destruction, ideas that are opposed to the tradition of legal deposit. However in the framework of legal deposit, some of the researchers asked why the BnF collects sites rather than asking producers to deposit them, as with printed material.

5. Potential services and information needs of researchers

Document the collection process

Given the problems regarding the use of the web in scientific work discussed above, the fact that a public institution such as the BnF is responsible for archiving the web is seen positively. The status of the library, the transparency of its procedures and the accessibility of its collections provide the researcher with a basic guarantee against the potentially anarchic and unverifiable nature of a web archive. Researchers questioned raised the idea that such an archive should allow them to cite a "call number" in referencing these archives; but beyond that a public institution should be able to answer the questions: How can these collections be characterised? Who made this archive? What is its status?

Individual archiving practices such as those already described do not allow a response to these questions, which makes their use especially problematic for a researcher, as there can be a personal bias in the collection that the use of a publicly-created archive should be able to reduce. The most important thing for a researcher is that the collection policy is clearly defined and communicated. The selection criteria must themselves be archived, as they may evolve over time. The institutions of the state are considered positively as preservation bodies, with the benefit of transparency. (One researcher notes the irony of the fact that the state is now responsible for preserving audiovisual documents created by leftists and anarchists in the 1960s and 1970s, when the anarchists themselves had little interest in preserving for posterity: "creative things, those without boundaries, tend not to develop archives".)

Describing the archive

Linked to this is the need for description and structure to allow researchers to discriminate between the sites collected. One researcher raises the problem that the web archive risks putting everything on the same level, as the web can give the impression that there is no differentiation, when in fact sites are structured both internally and externally by links.

Establishing a typology of sites is difficult; the researchers questioned tried to list criteria (by type of producer, by type of site...) and rapidly realised that the lists were infinite or that the categories overlapped. It therefore becomes clear that there is a risk in trying to apply traditional classification practices to a collection that does not apply traditional structures. The reactions from the interviews show the need to find new ways of talking about the archives; one researcher used metaphors from natural history as a way of creating classification based on the relations between sites.

At the level of each site, one researcher listed the information needed:

- the URL and the date collected
- the location of a page within a site and its structure
- the sites that link to that URL and the sites to which it has links; what interests the researcher as much as the site in itself is its place within a network of sites.
- usage statistics, such as the number of visitors, of views, and its ranking in Google results pages. To allow researchers in the future to differentiate between sites, the popularity and reputation of a site should be shown in some way. However this is not in terms of absolute numbers of audience, as the Internet consists of "niches" where a site may be important to a small audience.

Cooperation and communities

While it is a truism that the web is infinite and no-one can know it all, this is not necessarily a view shared by everyone, and many students and researchers have the impression that they know the web well in their specific area. The practices used to monitor the web, described above, can create "egocentric" networks which can filter out huge parts of the web.

In addition, there are dark or hidden areas of the web, sites that are intentionally not referenced in the usual way to stay discreetly hidden, particularly in the case of underground art, for example. For this reason, the researchers questioned were doubtful about using a group of experts to identify important sites, with one proposing instead asking a larger number of people for a few sites that they follow, and renewing the questionnaire every six months. Rather than trying to judge which sites are important, this approach would identify the sites that people use.

In relation to focused collection of the web, one researcher immediately mentioned "research programmes", underlining the fact that there is no simple way to crawl the web, that any collection requires long-term, collective exploration and observation. This work in advance of the collect should also allow the creation of tools allowing the researchers to navigate among the sites collected. The idea of "research communities" was also raised, whereby researchers could share sources and information; in the opinion of one researcher this is necessary if the archives are going to be exploited fully.

Another researcher working on net art points out that this kind of material is difficult to find and search engines are not much use, and that manual indexing would therefore be necessary to make the collections usable. This work, seen as collective and passing by "portals" and "research communities", would allow qualitative description and organisation of resources. As this work cannot be done automatically, the BnF should try to capture the work that is done by researchers using its holdings. While this was related to traditional academic practices, one researcher remarks that the knowledge that amateurs may have of specific areas of the web should not be ignored, but should also be exploited and preserved.

6. Conclusion and perspectives

While this study was based on a limited number of interview subjects, their responses provide many elements which will be helpful to the BnF in planning the future development of its Internet legal deposit.

Regarding **content** and **selection policy**, the researchers all agreed that it is impossible to predict what material will interest professional or amateur researchers in the future; but also that some degree of selection is legitimate given the volume of data that exists. In this, the study confirms that the approach chosen by the BnF, of a "mixed model" combining large-scale crawls with focused crawls based on manual selections, seems to respond best to this contradictory demand. This policy should therefore be maintained, however the crawls should be made more reactive to the evolution of the web: keep track of nodes and networks, the most popular sites, Google results. The most important is to preserve not only isolated elements but the activities on the web that show new trends, be they social or commercial.

Regarding **services** and **promotion**, the decisions and criteria that are used in the selection policy must be justified, documented and made visible. Tools should be put in place that allow researchers to know whether a site has been archived and especially to be able to find their way in the archives and discriminate between sites. As web archiving is a new concept to most researchers, to improve communication and promotion it is necessary to explore different metaphors that will allow different user groups to represent and imagine the web archives.

Finally, the role of **communities** and **cooperation** is essential. This includes not only communication for researchers via participation at conferences or other bodies and projects, but to engage with researchers and amateurs working with the web and to involve them in the creation of methodology and identification of sources within the archives, that may be used in access tools. Involving the research community directly should help not only to increase the "legitimacy" of web archives in scientific work but also to encourage researchers to make use of the resources contained in web archives.

Notes

[1] The responsibility for Internet legal deposit is shared between the BnF and the National Audiovisual Institute (INA). For more details on legal deposit law and its implementation see Peter Stirling, *et al.*, "The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future," in *Proceedings of the 77th IFLA General Conference and Assembly, Puerto Rico, 2011*, (IFLA, 2011).

[2] Sara Aubry, "Introducing Web Archives as a New Library Service: the Experience of the National Library of France," *Liber Quarterly*, 20(2), 2010.

[3] See for example Fabienne Greffet (ed.), *Continuerlalutte.com: les parties politiques sur le web* (Paris: Presses de la Fondation nationale des sciences politiques, 2011).

[4] The complete study in French is to be found on the BnF website : Philippe Chevallier and Gildas Illien, *Les archives de l'Internet: une etude prospective sur les representations et les attentes des utilisateurs potentials*, (Bibliothèque nationale de France, 2011).

[5] Ariel Bleicher, "A memory of webs past," in *IEEE Spectrum* (March 2011).

[6] See for example John A. Bargh, Katelyn Y. A. McKenna, "The Internet and Social Life," *Annual Review of Psychology*, 55 (2004) : 573-590; Antonio Casilli, *Les liaisons numériques : vers une nouvelle sociabilité?*, (Paris: Seuil, 2010) ; Jean-Claude Kaufman, *Sex@mour*, (Paris: Armand Colin, 2010).

[7] This is particularly true for sites hosting illegal content, but also sites whose content is personal or potentially sensitive. One researcher cited the case of jennicam.com, a site set up by an American student in 1996 which broadcast a webcam installed in her bedroom 24 hours a day. The site closed after seven years in existence, and today only a few screenshots remain.

[8] For example, the "Sida-Mémoires" archive, stored at IMEC, includes screenshots of sites and personal pages relating to the history of AIDS, where the Internet has played a central role. See Institut Mémoires de l'édition contemporaine (IMEC).

[9] Internet Archive.

[10] See also the report by the Oxford Internet Institute, commissioned by the International Internet Preservation Consortium (IIPC). Eric T. Meyer, Arthur Tomas and Ralph Schroeder (Oxford Internet Institute, University of Oxford), *Web archives: the future(s)*, (IIPC, 2011).

[11] ASCII art (dating from the 1960s-1980s) involves creating images solely using letters and other characters available in ASCII coding.

[12] By law, the access to the BnF Internet archives is limited to accredited researchers and strict conditions are defined regarding the use of the archives. For more details see Stirling et al. "The state of e-legal deposit in France...", pp. 19-24; and the BnF website.

[13] France Lasfargues, Clément Oury and Bert Wendland, "Legal deposit of the French Web: harvesting strategies for a national domain," in *Proceedings of the 8th International Web Archiving Workshop Aarhus, Denmark, 18th & 19th September 2008*. See also Michaela Mayr, *International Internet Preservation Consortium Harvesting Practices Report* (IIPC, 2011).

Bibliography

[1] Aubry, Sara. "Introducing Web Archives as a New Library Service: the Experience of the National Library of France." *Liber Quarterly* 20 (2010).

[2] Bargh, John A. and McKenna, Katelyn Y. A. "The Internet and Social Life," *Annual Review of Psychology* 55 (2004): 573-590.

[3] Bibliothèque nationale de France. "Digital legal deposit: four questions about Web Archiving at the BnF".

[4] Bleicher, Ariel. "A memory of webs past." *IEEE Spectrum* (March 2011).

[5] Casilli, Antonio. *Les liaisons numèriques: vers une nouvelle sociabilitè?* Paris: Seuil, 2010.

[6] Chevallier, Philippe and Illien, Gildas. *Les archives de l'Internet: une etude prospective sur les representations et les attentes des utilisateurs potentials*. Bibliothéque nationale de France, 2011.

[7] Greffet, Fabienne (ed.). *Continuerlalutte.com: les parties politiques sur le web*. Paris: Presses de la Fondation nationale des sciences politiques, 2011.

[8] Institut Mémoires de l'édition contemporaine (IMEC). "Sida-Mémoires."

[9] Internet Archive.

[10] Kaufman, Jean-Claude. *Sex@mour*. Paris: Armand Colin, 2010.

[11] Lasfargues, France; Oury, Clément and Wendland, Bert. "Legal deposit of the French Web: harvesting strategies for a national domain." In *Proceedings of the 8th International Web Archiving Workshop, Aarhus, Denmark, 18th & 19th September 2008*.

[12] Mayr, Michaela. International Internet Preservation Consortium Harvesting Practices Report. IIPC, 2011.

[13] Meyer, Eric T.; Tomas, Arthur and Schroeder, Ralph (Oxford Internet Institute, University of Oxford), *Web archives: the future(s)*. IIPC, 2011.

[14] Stirling, Peter; Illien, Gildas; Sanz, Pascal and Sepetjan, Sophie. "The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future." In *Proceedings of the 77th IFLA General Conference and Assembly, Puerto Rico, 2011*. IFLA, 2011.

About the Authors

Peter Stirling is a digital curator in the digital legal deposit team at the Bibliothèque nationale de France (BnF). He works on services for users of the web archives, as well as day-to-day web archiving activity and the international activity of the team in the context of the International Internet Preservation Consortium. He holds an M.A. in English Literature and an M.Sc. in Information and Library Studies, and previously worked for an online information portal for health professionals in the UK and in online information monitoring for the French National Cancer Institute before joining the BnF in 2009.

Philippe Chevallier holds a PhD in Philosophy from the Université Paris-Est and a degree in Mathematics. He joined the Bibliothèque nationale de France (BnF) in 2008, and is currently project manager at the strategy and research delegation, responsible for the planning of studies. In the area of audience studies, he has recently conducted a survey, with Laure Rioust and Laurent Bouvier-Ajam, to be published as "Consultation of manuscripts online: a qualitative study of three potential user categories" (Digital Medievalist, 8, 2012, forthcoming).

Gildas Illien is director of the bibliographical and digital information department at the Bibliothèque nationale de France (BnF). After 6 years devoted to the implementation of web archiving at the BnF, he is now in charge of pushing the Library's metadata and catalogues towards the web of data in the Linking Open Data (LOD) environment. He has also served as Program Officer and Treasurer of the International Internet Preservation Consortium, representing this organisation in international conferences and organizing international cooperation for software development, advocacy and collection building in the field of web archiving among 40 heritage and research institutions on three continents (Europe, North America, Asia). A digital curator with academic background in management, political science and sociology, he has published several articles and book chapters about web archiving (collection development, usages, international cooperation). (*Photograph of Gildas Illien by Didier Pruvot.*)