Date submitted: 02/06/2009

Metrics and Strategies for Web Heritage Management and Preservation

BERMES, Emmanuelle;
ILLIEN, Gildas
Bibliothèque nationale de France
Paris, France

Meeting: **92. Statistics and Evaluation, Information Technology** and **Preservation and Conservation**

WORLD LIBRARY AND INFORMATION CONGRESS: 75TH IFLA GENERAL CONFERENCE AND COUNCIL
23-27 August 2009, Milan, Italy
http://www.ifla.org/annual-conference/ifla75/index.htm

I. Introduction: digital heritage isn't inherited

The acknowledged goal of conservation is to protect heritage collections from the injuries of time in order to keep them accessible. However, there is no such thing as an obvious definition of "digital heritage". Over the centuries, many heritage institutions seem to have looked at their collections as an accumulation from the past, which would only require special conservation and care after a certain time. Conservation seems to start or be taken seriously into consideration only when damage or explicit risk occurs. It is therefore often seen as a specific branch of the collection management workflow, which usually comes at the end of a wider and longer process (acquisition, description, usage…) and requires dedicated expertise.

The massive development of cultural artifacts produced in digital format – either digitized or born digital – involves a critical change of this paradigm. It is practically impossible to save or restore digital collections after the data has been corrupted. Digital documents do not usually decay over periods of time but disappear at once, in one single shot. This is why we need to look at their conservation as a complete process starting from the moment they are created. This is also why it is preferable to talk about *preservation* rather than conservation when dealing with digital material: preservation needs to be taken into account from the very beginning and at every step of the document life cycle.

Moreover, we lack distance and perspective – time – to know which segments of contemporary and ever-growing digital production should or should not be part of the scope of national or local heritage. With hindsight, thinking back about the legal deposit of prints, we might consider that it was relatively easy neither to select nor sample from global production: we could just take everything published. The change of scale now requires making more radical selections as the value of some things to be preserved rather than others isn’t as obvious as it might seem at first glance. From that perspective, it could be argued that the digital librarian faces the same challenges as those of the archivist. Who is he to judge and to choose what will be of interest to future generations? Value cannot be predicted. History has demonstrated that many items, such as popular literature (e.g. "bibliothèque bleue"), engravings used as advertisements, political flyers, and similar objects considered worthless or of no interest at the time they were created, ended up being most valuable for contemporary curators and historians. Historians have a natural propensity to challenge all curators’ plans and to make them fail: they are constantly searching for what was not expected to be searched. Their goal is indeed to gather original sources representative of the culture of a time or era. For instance, it would be absurd to study the life and habits of teenagers in the early years of the 21st century without looking at samples of their video games, blogs and MP3s. These digital items are part of today's heritage, in that they reflect contemporary culture. They will eventually end up being part of our collective history.

When it comes to the web, we cannot define heritage as material inherited from the past: this heritage is still too young to be inherited. It is a wide and living collection to be approached as a world or a city constantly under construction. It is more a network, and a process – *history in progress* – than a stack of items. At every stage of digital collection management, similar questions to those which used to be addressed from a strict conservation perspective need to be raised: whether you want to select, describe, highlight or save specific contents within the collection, you need consistent terms of reference in order to analyze this information and decide its value all the way through. As a result, we may consider that preservation, as a concept and a set of actions, embraces the whole workflow of digital collections.

This presentation deals with the management of these very collections with a focus on evaluation and metrics for web archive preservation. As ISO is launching this year a new topic on "Statistics and quality issues for web archiving", together with a working group to produce a technical report on these issues, our goal is to contribute to this discussion by highlighting best practices observed in web archiving communities around the world, along with some suggestions of our own. [1]

Our first assumption is that the notion of value (whether defined by cost, content, usage, scarcity or risk) is a key to any preservation decision and planning, and thus requires a set of objective standards which are much needed in order to qualify, characterize and measure the collection value and take action accordingly at every step of the workflow.

Our second assumption is that web archive metrics, like other digital collection preservation metrics, need to be designed for business relationships. We will illustrate this approach with the OAIS model (Open Archival Information System, ISO 14721), which defines a framework to organize roles and actions within a digital preservation environment.

We chose to focus on the example of web archives, which can easily be expanded to other forms of digital material, though it does have some specific patterns. Web archives are an interesting

[1] This paper builds over previous work on web archive metrics presented at IFLA conference in Québec. See: Gildas Illien, « L’archivage d’internet, un défi pour les décideurs et les bibliothécaires : scénarios d’organisation et d’évaluation ; l’expérience du consortium IIPC et de la BnF », proceedings of 74th IFLA Conference (Québec, Canada), 2008. www.ifla.org/IV/ifla74/papers/107-Illien-fr.pdf .

example for a digital collection because they are an emerging field with few standards in general, and no official standard at all when it comes to statistics. Moreover, among the patterns and requirements of web archiving management, the most striking ones, such as their distinct scalability challenges, are very likely to become an issue for all types of digital collections.

II. The challenges of web archive measurement

Web archiving has been an increasing concern for heritage institutions since the end of the 90's. By then it became clear that the web was to play a major role in the development of communications and the dissemination of knowledge. To many libraries and library users, the Internet quickly proved to be an essential tool for information research and retrieval. The next step was to acknowledge the fact that, as a network and a publishing space, the web was likely to become the virtual and fragile repository of World knowledge. Due to the rapid growth and the diversity of web contents, new techniques based on automatic harvesting and "crawler" or "spider" robots were invented in order to capture, index and search this new part of our heritage[2]. Many national libraries or archives were mandated by their governments to endorse this new task, often as an extension to existing legal deposit regulations. Whether performed selectively and requiring permission from web publishers (selective harvesting) or on a larger scale, for comprehensive national domain crawls (bulk harvesting), web archiving is still an experimental program in some parts of the world, while becoming a full production activity in others.

Many of the resources we find on the web are somewhat familiar to a librarian and fall into general or special collection categories. Even a large, encyclopedic library such as the Bibliothèque nationale de France (BnF) can find online avatars and replicates of most of its traditional document types. This is not only because the process of mass digitization started a few years ago, but also because many born digital materials primarily derive from widespread cultural objects and artifacts: e-books, e-journals, Government publications, photographs, maps, drawings, scores, ephemera, films, music, advertising, movies, games, etc. Some online documents could even qualify as e-manuscripts. There is indeed a strong continuity of library material and usage in the electronic Age. We might be facing a screen, but we still read books and newspapers, we still watch movies, play games and listen to music.

Precisely because most collections have a web equivalent, librarians experience the temptation to project or simply adjust existing policies and classification schemes when they start looking at web archives in order to define rules and criteria to characterize and organize them. Couldn't we just approach web archiving measurement with the same categories which have shaped library collections in the past?

In fact, the web is not restricted to the familiar. On the contrary, it tends to extend and expand beyond our conception of material collections. Furthermore, the technologies and organizations invented to produce, index and preserve web archives are quite different from those used for non

[2] The Californian, not-for-profit foundation Internet Archive acted as pioneer and key developer in this field as of 1996. It was later followed by a group of national libraries from Europe, North America and Australia. Together they created in 2003 the International Internet Preservation Consortium (IIPC) which now gathers 39 institutions on four continents. Most of them are national libraries or National Archives which have joined their efforts in order to raise awareness in this field and develop collaboratively open source software aimed at harvesting, rendering and preserving Web collections. See: www.archive.org and www.netpreserve.org .

digital collections. Most of the metrics used in international standards for library measurements therefore do not apply to web archiving. The web has a few distinctive patterns which must be outlined in order to understand why we need to think differently and creatively.

Scalability: the web is big.

Even in comparison with the largest library collections, the web is extremely big. Looking at the figures from the past decade, its growth seems hyperbolic and unpredictable[3]. As a result of this critical change in scale and scope, many of the collection development criteria and processes based on a per-item analysis (such as cataloguing or author/publisher identification), and which usually serve for measurement, can no longer be used because they are simply too expansive and therefore not scalable.

Internationalization: the web is global.

Scoping the limits of a national web domain (for legal deposit purposes for instance) is one of the greatest challenges for national libraries. Websites cannot easily be localized on a specific territory. Technically, legally, culturally, the very notions of nation and territory are very difficult to project on the web, which is essentially an international medium. As to language discrimination, it is also seriously challenged. Although automated language recognition allows for harvesting robots to take languages as a selection setting, this proves valuable only to countries whose language isn't spoken outside their borders. Hence the measurement categories attached to territory, language or geography can hardly apply.

Speed: the web is fast.

Even if some websites have been online for over a decade, the rate of change and destruction for an online publication is definitely higher than for printed material, especially in areas such as politics, news media, or when a particular event takes place (for example: September 11) [4]. The formats found on the web also change very fast. As a result, collection and preservation strategies need to be extremely reactive and fast, which leaves little time for detailed analysis and measurement before and during harvesting. This involves being ready to save the ephemera, the unscheduled - and precisely to plan collection development and preservation whilst taking into account a high level of risk and uncertainty as a rule rather than as an exception.

[3] Recent statistics illustrate the issue for the French context: in 2008, ca. 70 000 new titles of books were collected and catalogued by BnF legal deposit services; at the same time, the French Domain Name Industry Report stated that the .fr Top Level Domain then consisted of over 1.3 million domain names, while another report from AFNIC, the French Registrar for .fr, evaluated that less than 30% of French websites were actually registered under a .fr domain. See: Afnic, *Observatoire 2007 du marché des noms de domaine en France*,Saint-Quentin-en-Yvelines, Afnic, 2007, 96 p. www.afnic.fr/data/actu/public/2007/afnic-observatoire-domaines-france-2007 .

[4] According to a report from the PLANETS project, the average life cycle for a website lasts 44 days. Cf: http://www.smh.com.au/articles/2003/10/16/1065917549444.html

Virtuality and multiplicity of document types: the web is intangible and diverse.

Libraries are used to handling a wide variety of document types and many Library statistics are based on the distinction between physical media, e.g. books, periodicals, discs, etc. However, the web consists of virtual files with no physical existence to which such statistics could be applied. Furthermore, web pages display extremely heterogeneous data files, formats and players (plug-ins), many of which are neither standardized, documented nor open-source, which makes them even more challenging to identify, characterize and hence to preserve in the long run[5]. Both the diversity and the rapid change of formats and publication types thus make it very difficult – if ever possible – to use existing standards or to rely on systematic and scalable format identification and characterization tools.

Twilight zones: the web is for everybody and everything.

Almost anyone can publish almost anything almost anywhere on the web. Technology has made it possible for every end user to become a writer and a publisher. Many web documents (such as blogs and wikis) have hundreds of contributors. Social networks and virtual worlds such as Facebook or Second Life have created what we call twilight zones because the status of these arenas remains unclear both to their visitors and, sometimes, to the contributors themselves. Hence, the Internet has drastically changed the lines of the public space, developing many of these grey zones which are difficult to qualify as publications according to usual library terms of references. Legally, in the French context at least, any website falling into the national domain scope is regarded as a publication and is therefore subject to legal deposit. But ethical and collection issues obviously arise when it comes to harvesting and accessing this data. How far should the Library go? What is to be included or excluded from national heritage?

Web document structure: the web is a puzzle.

The web is by nature a composite medium, made of a multiplicity of files assembled following a uniform and standardized structure. Granularity of web pages or websites, and hypertext linking between them, make it very difficult to delineate what librarians would call a document, or a homogeneous intellectual entity (by the same author(s), with related contents, from the same period, etc.) The physical structure of the web, based on digital files and domain names, creates a publication space where information is structured as a network, and this physical structure rarely matches the logical or intellectual structure of websites. For a library, capturing the web involves sampling the whole e-publication process, i.e. its organic structure and its links and not only its data. As a result, web archiving means changing our vision of what a document is, and even if we manage to do this, it remains difficult to define and to name the limits of this new document. When it comes to collection metrics, what should we then actually count? Which pieces of the puzzle should be counted as unique "documents"?

[5] Reports from a recent BnF domain crawl from 2007 showed for example that this collection included a total of 1 604 different MIME types and that, in comparison with similar reports from the two previous years, very significant changes were observed both in the MIME type list and in its distribution.

III. Metrics scenarios for web archives

The metrics we propose are either already being used by several institutions in the world (e.g. some national libraries, national archives, foundations such as The Internet Archive) or more creative proposals based on BnF's current work and thinking for its own web archiving workflow[6]. Before listing them, we will propose a general framework to approach these statistics from the perspective of organization. Digital collection management indeed comes together with a drastic change in work relationships. We believe it is of the utmost importance to acknowledge this situation as it can make a true difference when choosing some statistics rather than others: to be usable and useful, statistics must serve practical needs and be designed for the specific communities who will use them.

Web archiving and preservation communities: a new distribution of roles

The OAIS model provides a useful framework for defining roles and actions in a digital preservation environment. This conceptual model can be used to define the system aimed at preserving digital objects, but also the organizations and workflows designed to achieve this task. A typical web archiving workflow involves three types of people:

- Curators, also known as digital librarians or web archivists, have responsibility for collection development and policy. They are to take the decisions to include, exclude – choose – items or sets of items at every stage of the workflow, from acquisition to preservation: they "feed" the workflow with seeds (addresses of websites to be harvested) and decisions. They are the ones who need to assess the value of things in order to take heritage decisions everyday and in the longer term. In the OAIS model, this category includes "Producers" – those who create or harvest the digital objects– and preservation experts, the latter being part of the functional entity called "Preservation Planning".

- Technicians – crawl operators, computer scientists, engineers, etc. – run operations on the other hand: their task is to develop, build, maintain and monitor the workflow. Though institutions have started recruiting and training staff with mixed expertise, the technical environment in which web harvesting and computing takes place usually requires a strict distinction between those taking decision and those taking action. This distinction is well defined in the OAIS model, the technical staff being represented by the "Administration" functional entity. Administration is in charge of the digital Archive on an operational and day-to-day basis.

- An additional layer of decision-making is usually needed. We call project **managers** or digital library managers those who watch over the resources and the framework. The OAIS model also refers to them as "Management" and considers them as an external player, closely linked to Administration. Their focus is to analyze cost, risk and policy at a higher level, and to make sure librarians and technicians work together efficiently. Their job is also to look after funding, lobbying, negotiations and advocacy for digital collections.

[6] Reports from a recent BnF domain crawl from 2007 showed for example that this collection included a total of 1 604 different MIME types and that, in comparison with similar reports from the two previous years, very significant changes were observed both in the MIME type list and in its distribution.

The relationships between these three categories of staff tend to be much tighter and more frequent than in the past. For physical collection conservation, librarians are rather independent on a day-to-day basis: they require technical assistance only when specific actions are to be undertaken, e.g. restoration, deacidification, deinfestation, or when an accident requires extraordinary measures to protect the collections. In digital times, it is the other way around: as the material is stored on digital media, it is the role of technicians – Preservation Administration – to watch over the day-to-day workflow. And it is the technicians who require the librarians' – Preservation Planning – decisions for important actions such as launching a major ingest, indexing or migration task. Librarians and technicians are, in many ways, engaged in the process of swapping their roles. The librarian is giving up many of his daily tasks to the technician, who monitors their automation and takes responsibility for them on the librarian's behalf. In the meantime, managers are expected to act as "honest brokers" of this game.

The balance between these roles and skills underlies the digital preservation organization within the OAIS model. The "Administration" entity stands for technical expertise and decision making on a day-to-day basis. It is mandated to apply the policy defined and utilize the resources allocated by the "Management". The "Preservation Planning" entity watches over external technology changes and monitors the efficiency and the risks of the internal system on behalf of target user communities. This entity must therefore embrace the expertise of collection managers along with the end-users' expectations and interests. Its contribution is crucial to define, prioritize and test preservation operations. However, it doesn't take any decision nor play any part when it comes to running these operations. The goal of this distribution of tasks is to facilitate a fair balance between the inertia which is often characteristic of I.T. Administration and the more unstable dynamics attached to fast technology changes in digital hardware and software environments. Good communications between these entities, based on internal contracts (called "service level agreements") are thus crucial to this balance.

Metrics for digital collections management therefore need to be designed for business relationships. Our assumption is that the three communities need common standards to manage the collection together. However, specific types of metrics can also be useful in light of their dedicated duties and concerns.

A- Sites and seeds: metrics for collection development

Digital curators are primarily involved at the very beginning and at the very end of the web archiving workflow: selecting websites, giving priorities to some of them (by allocating a dedicated budget and deciding how deep and how often they should be harvested), contributing to the supervision of their acquisition (quality assurance), and ultimately promoting the collections and helping researchers using them, either online or in reading rooms. In some institutions, curators also catalogue harvested websites but we won't cover this case here. As preservation experts, digital curators may also be involved in long term preservation decisions but this will be discussed later on.

Considering these tasks, their measurement requirements for collection development are typically those of a collection planner/developer. Curators first need metrics to organize the allocation of their acquisition budget (measured in number of URL, workforce or computing effort) and to evaluate the quality of the service provided by I.T. Administration to collect the resources they ordered. They also ask for reports and metrics to evaluate the content of the collection after harvesting: they need to know what the collection is made of (characterization) and where it comes from (provenance). This part is all the more challenging given that the collection is big.

Seeds and orders

In order to define a service level agreement with the I.T. Administration, curators must be able to express and measure their expectations and objectives in terms which comply with harvesting technologies: their collection **targets** will become collection **orders** for the I.T. Administration. The most important concept is the notion of ***seed*.** A seed is a URL (a website or a piece of a website) which comes together with settings which are required to express scope and priority of the order. A seed can thus refer to a whole domain name (bnf.fr), a host part of this domain (only some files stored on a dedicated server), or even a smaller segment of the website (e.g. a page, or even a PDF publication hosted on a page). Curators must also specify how often or at which exact date the resource is to be harvested (e.g. everyday, every month, once a year, only once, at a given time, etc.).

Orders can be expressed in different ways depending on the harvesting scope and scale. For selective harvesting, where all websites are somehow "curated", it is possible to process and to check such information at the level of every website, hence relying on a per-item approach. The information about the seeds (input) can then be compared to the information about the harvested collection (output), measured in number of files or size of collection. For example, checking and comparing over time the number of harvested files for a given and recurrent order will help evaluating the quality and the depth of the crawl. This information is also valuable to analyze whether I.T. Administration answered the initial order, and whether the curators expressed their objectives properly. Such metrics will prove very useful when organizing further negotiations between both categories of staff as part of the service level agreement discussion.

Jobs and collections

It is very difficult to manage collections at site level when production reaches a certain volume of activity. Orders and seeds then need to be grouped in bigger "packages". There are different approaches to this. From the IT Administration and the monitoring point of view, the most practical approach is to group seeds in what we call ***jobs***. A job is a list or a "queue" of seeds to be harvested together as a consistent work package. One may decide to group seeds in the same job because they belong to the same order (for instance: seeds ordered by a specific collection unit of the Library) or because they are of similar size (grouping together big websites or websites to be crawled at domain level will prevent them from "blocking" the harvesting of smaller websites which can then be processed much faster, separately). One may also choose to group seeds which need to be selected very fast at the same moment (e.g. during an election or right after a natural disaster). Counting jobs and related metrics (e.g. average duration or average size of harvested data for a given, recurrent job) proves extremely useful in evaluating acquisition costs. Such information will be used by digital curators and I.T. Administration to plan together harvesting activities and estimate required resources over a period of time.

Though there is no agreement on best practices in this area, grouping seeds with the same provenance or serving a similar scientific purpose, also leads to building what is often called ***collections***. This approach is end-user oriented and sometimes referred to as "resource discovery". As it is not easy to search within the web archives (few institutions provide a comprehensive catalogue or a full text search engine), libraries are currently exploring other ways to highlight special parts of their born-digital resources in order to promote them as "keys" to search the global, usually massive, collection. For instance, the Library of Congress groups its web archives by collections such as *September 11*, *Katrina* or the *Tsunami*. Similarly, BnF recently launched a dedicated interface to access a sample of its collections related to *e-diaries and digital lives* and another one about *Web political campaigns since 2002*. It is hard to predict how such approaches will develop in regard with cataloguing and indexing evolutions, but it is worth acknowledging that between the website level and the mass level, there are middle layers or levels of granularity to explore. Specific metrics will be needed for managing collections at that level.

For bulk, exploratory harvesting, a high-level approach is more appropriate and inevitable: we then only look at the total number of URLs in the initial seed list and at the key settings of the domain crawl. In that case, a complete domain crawl covering, for instance, a national domain for a period of 5 to 6 weeks will be considered a collection of its own: the French domain crawl for Year 2008, for instance. Reports produced by harvesting robots provide statistics such as the size or the total collection, measured in number of documents (files) or by size (Gigabytes or Terabytes), the distribution of documents by top level domains (such as .fr, .de, .com or .org) and by format types (cf. infra). This information can be compared from crawl to crawl, and between crawls, and analyzed in many ways to make policy decisions for future harvesting campaigns[7].

B- Files and bytes: Metrics for collection processing

Metrics for web archive processing are intended to facilitate the technical steps of collection management: ingesting into the archiving environment and preservation strategies such as migration or emulation. These are more technical and, in a way, less specific metrics than those we proposed for collection development.

The preoccupations of technicians (who supervise harvesting, indexing, storage and preservation operations) are mainly about the number, the diversity and the size of the flows and the file storage associated with these operations. Beyond the amount of data, which is still the most useful indicator for identifying critical points in the workflow, I.T. specialists also use indicators concerning **production, workload and performance.** Although these are not specific to archiving, they enable the suitability of an I.T. infrastructure to be evaluated, bottlenecks to be spotted and also a certain visibility regarding the cost of the service provided.

In the **preservation phase**, the dialogue tools between librarians and technicians are the indicators which record the events in the life cycle of digital documents in the archiving system (events recorded as "provenance information") and the audits to check the integrity of archived

[7] See, for instance, BnF's comparative analysis of the 2007 and 2008 French domain crawls. France Lasfargues; Clément Oury ; Bert Wendland, « Legal depositof the French Web : harvesting strategies for a national domain », proceedings of the 8th IWAW Conference, International Web Archiving Workshop (Åarhus, Danemark), 2008. http://iwaw.net/08/IWAW2008-Lasfargues.pdf

information packages (fixity, checksums). They ensure that digital objects are not degraded in time or during preservation operations such as media refreshment.

Nevertheless, such operations can only be considered or managed in a satisfactory way if we have a clear vision of the entire collection, both quantitative and qualitative. This vision must be based on objective elements, which will not be susceptible to ambiguity or misunderstanding when discussions take place between technicians and librarians. These elements can be considered the **core elements** - agreed indicators, necessary for a perception of the collection shared by all.

The number of files

The web is a vast system of addresses and requests, allowing the storage, the display and the linking of files held on servers; nothing more, nothing less. This system limits us to the atomic unit of the file - the URI, the lowest common denominator of the web and its archives. At first sight the results seem unusable because the number always appears too high compared to traditional collections. Any library is, however, confronted with the impossible task of aggregating very different documentary units, whether they be digital or not. When a library must take account of its acquisitions, isn't it the same difficulty? BnF holds a wide diversity of objects: millions of books, naturally (which we count in titles or copies) and millions of periodicals (titles both alive and dead, and which are counted in issues or bound volumes), but also audio-visual media and a great number of specialized collections whose unit of measurement varies each time. What is there in common between a lithograph, a photograph, a medal, a map, a globe, an anthology (of tracts, of ephemera, etc.), and a collection of Sarah Bernhardt's stage outfits?

It is exactly the same for digital collections likely to be kept in the long term, whether they are created digitally or the result of a digitization process. Most of the time these digital objects are complex entities, of which the file, in the I.T. sense of the word, is just the smallest component; the lowest granular level, but rarely a coherent object from an intellectual point of view. The aggregation of files is above all convenient, and we can use it while admitting that these figures cover very different types of documents and contain a number of duplicates. The number of files collected serves to measure the development of the activity with time, and to make comparisons. The essential remains, as for any statistics, that the measurement is uniform and sustainable.

The size (TB- GB)

The other idea is to look for a solution, once again, among existing practices. What does a library do when it must make the case for a project extending, relocating or reconditioning its collection? It measures all the diversity of its collections in a simple, unified way - the linear meter or foot, in order to evaluate the costs and organize its operations. The same method can be applied to web archives, but by replacing physical storage needs (in the stacks) with digital storage methods where it is the size of the data that is the benchmark. At a time when many institutions are committing to ambitious digital repositories and the cost of digital storage has become a crucial budgetary issue, this approach seems all the more relevant. It is thus in gigabytes (for the institutions with selective methods) or in terabytes (for the more greedy) or even petabytes (tomorrow) that web archives can be measured equally.

Counting and sorting by MIME types.

Reports provided by harvesting robots produce statistics relative to the MIME type of the files acquired, e.g. 70% of files text/html, 15% of files image/jpeg, 5% of files image/gif, etc.[8] This is very valuable information from a preservation point of view, because it is knowledge of the types of files present in the collection which will enable us to assess the risks and the priorities and to define migration or emulation strategies for long term data archiving.

However, we know that the MIME types, as they are generated at the moment of harvesting, are often peppered with errors. Having acknowledged that the accuracy of this information is questionable, preservation experts envisage putting in place automatic procedures to systematically identify and qualify files, enabling us to generate more reliable file-type information at the moment of ingest. This question of identification remains a challenge today, first because it requires the setting up of international format registries, a process which is already underway, but far from completed[9]; and second because the resources necessary for characterizing all archived web files are enormous.

So should we reject MIME types as indicators for web archives? They remain the only objective means at our disposal for a broad brush assessment of the general distribution of a collection by file family. The information offers a documentary interest because it is a useful means of qualifying the distribution of the content harvested en masse (text, video, images, sound, etc.). We can therefore consider, even in the absence of a precise automatic identification process, that the MIME type is one of the major indicators of collection knowledge. It must, however, be treated with the appropriate caution.

(W)ARC files

Finally, many institutions use the (W)ARC file, which allows bulk storage of files as they are copied by the robot, as a complementary unit of measurement of their collections.[10] The (W)ARC format is a file container format initially designed for web archives, currently being standardized by the ISO. The use of (W)ARC to measure the size of a collection is open to debate, because it is a level of physical structure, created artificially, for practical reasons, at the moment of harvesting. It tells us nothing of the content itself or its structure. Nevertheless, this question comes back to the problem of managing the granularity in a digital repository, where the OAIS model recommends managing information in information packages whose size, and thus granularity, can vary according to whether it is the ingest stage (Submission Information Package or SIP), archiving stage (Archival Information Package or AIP) or diffusion stage (Dissemination Information Package or DIP).

[8] The registry of existing MIME types is maintained by the IANA at http://www.iana.org/assignments/media-types/.

[9] Cf. the projects to merge the GDFR and Pronom format registries as part of a new project called UDFR: The *Unified Digital Formats Registry*. See: http://www.gdfr.info/udfr.html

[10] When a file is harvested from the web, it is copied into an ARC container file (http://www.archive.org/web/researcher/ArcFileFormat.php), with the metadata collected during harvesting. ARC files can contain multiple digital objects: they are closed (i.e. the copying of files stops) only when they have reached their maximum size of 100 megabytes. So ARC files work like the digital equivalent of cardboard archive boxes that are filled with all sorts of documents. Today we talk about (W)ARC files because this format will develop towards the WARC format (http://bibnum.bnf.fr/WARC/index.html), which offers extended handling functions (e.g. the management of duplicates or of converted files). Its target size has been enlarged (1 gigabyte) to cope with the increasing size of objects on the Web. Its definitive standardization is expected from the ISO during 2009.

For all those managing large volumes, handling files one by one presents a real challenge, given the number of information packages to be taken into account. The (W)ARC format constitutes a unit of conservation and management which will be a reference throughout the life-cycle of a digitally created document. Quantifying using (W)ARC and cross referencing this information with statistics on the MIME type enables collections to be documented with a view to their preservation.

C- Metrics for collection management

Web preservation managers must be able to define collection value and preservation objectives at the high level, and to guarantee the best quality or security for the collection at the lowest cost. Comparing cost and value is therefore a crucial part of their job. The metrics they need to do so are also useful to report and negotiate at the higher, political and funding, top management levels. This is all the more important now that the issue of digital preservation technical and economic sustainability has become a critical topic for many heritage institutions and international organizations. [11]

How can we demonstrate and measure the cost of web archiving?

If the Library outsources the harvesting activity to a third party, then it is easy to know the cost of this service. On the other hand, estimating the costs of in-house harvesting is no more challenging than evaluating more familiar processes such as cataloguing. This estimation is based on four main types of expenses: hardware, direct costs, software, and labor.

Hardware includes acquisition and maintenance of the material necessary to harvest, index, ingest and store the data. **Direct costs** are for computing, for example power and network (bandwidth). As to **software**, national libraries often use free, open source software, many of which are developed and maintained by the IIPC consortium. [12] This solution is good for interoperability and standardization, it also avoids paying for major in-house developments or licenses to commercial companies. However integrating such tools (which are usually not plug-and-play applications) requires a significant expertise and a certain amount of a developer's working hours, especially at implementation and integration phase and for each new release. Moreover, institutions always have specific and additional requirements, whether functional (attached to local collection and service policies, to the internal organization, culture and language), or non functional (e.g. linked to the I.T. operating environment). These do require dedicated developments which will add to the bill.

[11] Cf: The *Blue Ribbon Task Force on Sustainable Digital Preservation and Access* by the National Science Foundation and the Andrew W. Mellon Foundation in partnership with the Library of Congress, JISC, CLIR, and NARA.

[12] Heritrix, crawler, http://crawler.archive.org

Wayback Machine, to browe and search by URL: http://archive-access.sourceforge.net/projects/wayback

NutchWAX, for full text indexing and search: http://archive-access.sourceforge.net/projects/nutch

WARC Tools, to manipulate WARC files: http://code.google.com/p/warc-tools/

Last and certainly not least, come **labor costs**, measured as usual in FTE (Full Time Equivalent) or men-days. These are usually split into the three categories of staff we are already acquainted with: curators for collection development and preservation planning, I.T. administration for development and operations and management. We know human labor can be the most expensive factor in many processes, especially as we need qualified or highly qualified staff for digital management: most repetitive and manual tasks being automated, human intervention is required only where expert decision-making is needed. So anybody taking part in digital collection management usually represents (or should) a relatively expensive workforce.

The **harvesting scope and policy model** (bulk, selective, or mixed) makes a large difference in cost analysis. Bulk harvesting is definitely more expensive in terms of computing and storage but it can be relatively cheap labor wise. In some institutions, such as the Nordic national libraries, a handful of people actually monitor some of the biggest repository of digital documents. On the contrary, selective harvesting and per-item curation turn out to be much more expensive because many librarians are involved. When BnF ran its 2007 Election harvesting project, it involved 23 curators and 2 engineers during 8 months. We calculated the average cost of curation and harvesting for a seed and came to the figure of € 51 per website. 90% of this cost actually came from human labor and site curation in particular. This figure is estimated to be 14 times higher than for a website harvested in bulk mode.

How can we demonstrate and measure the value of web archives?

Managers willing to start or to develop a web archiving program face many challenges when it comes to convincing their internal and external stakeholders of the value of this material. For example, they need to demonstrate that seemingly "bad" contents (e.g. pornography, commercials, blogs…) are part of the heritage scope and will add scientific value to the Library collections in the longer term. They also need to demonstrate that potentially "dangerous" contents (either unauthorized by regulation – e.g. racist or child pornography websites – or technically challenging – e.g. websites with viruses, robot traps, spam, domain squatting –), which the Library might harvest automatically represent a risk that is worth being taken. Overall, web archiving advocacy is about explaining that born digital documents are just as "good" for the Library and its researchers as the more familiar documents inherited from the printed Age – that "*Today's web sites are the tomorrow's research collections*".[13]

A first approach is to demonstrate the **scientific or heritage value** of a given collection by comparing its contents with more traditional collections. A rather efficient way to do so is actually to use examples, illustrations and narratives rather than metrics. One can for instance explain that harvesting Election websites is exactly the same task as collecting political flyers and programs during a campaign (and usually even brings back more material to the Library). But one can also use figures, for instance the number of files for a given format that is comparable to a media that decision makers know well: show the total number of PDF files and compare them with prints; show the total number of JPEG files and compare them with photographs; the same can apply for video, sound, etc. Naturally, the best strategy to address these issues depends on

[13] Quoted from Martha Anderson, Director of National preservation Program at the Library of Congress – IIPC General Assembly, Ottawa, May 2009.

the Library's specific collection policy, statement and traditions. It is usually easier to defend web archiving in an institution mandated for legal deposit than in an institution with a more restrictive and targeted acquisition policy.

Another approach is to use **scarcity and risk analysis.** One can, for instance, calculate the rate of loss for a given collection, i.e. show how much of a web archive collection is no longer online. In 2007, as we were launching the French Election harvesting project, we demonstrated that a third of the previous (2002) Presidential Election collection was no longer accessible online. This showed the necessity to harvest the event. One can similarly calculate the speed of loss (e.g. these websites disappeared in less than 5 years), which shows the necessity to take action fast. These indicators will prove even more meaningful in a few years when many more online publications will have totally disappeared and the Library might be the only institution to have saved a copy.

Usage is of course a very useful way to demonstrate value, though in long standing heritage institutions such as national libraries, usage isn't expected to come at once. We know it can take years – sometimes, centuries– before a collection item becomes valuable to researchers. Of course, Library stakeholders prefer seeing usage metrics demonstrating that their investment is immediately rewarded by users, but this is unlikely to happen before the collections grow older and disappears from the living web. For legal and technical reasons, few institutions actually provide public access to their web archives, and even fewer provide online access. This is currently the most critical weakness for web archive advocacy and the estimation of their value.

So far the few institutions which have tried to define usage metrics for web archives usually use standard web or library measurements for e-journals and documentation. The Internet Archive provides online access worldwide and demonstrates its popularity and public purpose by using the number of hits per second to their website (www.archive.org), i.e. between 150 and 200 hits per second in 2008. At BnF, where access is restricted to researchers in reading rooms on the Library premises, we use the following monthly indicators: number of visitors (how many registered users accessed the web archive during the month?); number of sessions (same indicator, but counting only the session lasting more than 5 minutes) ; total number of viewed pages. Those figures only make sense when compared to other electronic collection usage statistics and over time. We use them to demonstrate that there has been a positive trend in public demand and use of the web archives since we opened this new service in April 2008. Though the figures remain quite low, with a yearly average of 35 visits per month in 2008, we are mostly interested in demonstrating that this figure gets a little higher every month. There will probably be more sophisticated and appropriate ways to measure web archive usage in the future, but we first need more institutions to provide access and to run and compare use cases.

Lastly, one can also **demonstrate the value of web archiving by showing it is cheap** – or actually cheaper than managing physical collections. If there is enough confidence in Web harvesting, may choose to harvest certain documents instead of acquiring them on a physical medium. Regardless of the collection value, the idea is to demonstrate that harvesting is a valuable way to manage collections because it is an economic one. Good examples of resources which are expensive to manage physically are for instance local Government publications and daily newspapers. For these resources, the cost of harvesting is much cheaper than human acquisition, cataloguing and communication. However, this approach still seems questionable in the long term. We can prove that the harvesting and ingest cost is likely to be cheaper than the management of paper (especially if there is no cataloguing) but we cannot demonstrate that it will also be true for long term preservation.

D- Summary and Core metrics

Four elements (number of files/URIs, number of Giga/Terabytes, MIME types and number of (W)ARC files) could be considered the **core metrics** for technical monitoring and cost analysis at all levels of the workflow. We saw that they could also be useful to target and monitor collection acquisition to some extent. Thus they can be used by all stakeholders to characterize and analyze the collection, and will be the basis for establishment of more specific value indicators with the goal of managing costs. From a specific preservation perspective, it is still hard to predict which of these indicators will be most useful in the future, as this part of the workflow has not yet been fully implemented anywhere. However MIME types and (W)ARC files are likely to be promising candidates, though one should remain extremely careful when interpreting them.

In 2008, the IIPC consortium ran a survey among its 39 members and found that more than 90% of the member institutions used the numbers of Giga/Terabytes along with the number of files/URIs as the main units to count their collections.[14] These figures are never perfect[15], but they are usually false in a rather consistent and stable way which makes them truly useful when comparing big collections over time and between institutions. They allow for simple, efficient collection measurement provided one takes time to explain them to decision makers. The following chart summarizes most of the possibilities we have identified and listed in the current state of the art.

[14] A summary of the survey results is accessible on the IIPC public website: http://netpreserve.org/publications/reports.php?id=005

[15] The calculation and comparison of the total number of files can prove challenging because there is no agreement on whether duplicates should be counted or not, whilst there are usually many duplicates (used as back up or archived in parallel) in a web archive collection. Deduplication is an important issue in web archiving, either at harvesting or ingest level. As to the size of data, measured in bytes, it can also be biased by taking account or not (depending on workflows and institutions) file compression processes on the one hand and the indexes on the other hand.

Measurement/ Indicator	Examples	To whom and for what are they useful?
A- Collection development		
Number of seeds (URI)	370 (selective) 550 000 (bulk)	Librarians, collection development: characterize collection target and order, define and measure collection objectives, evaluate selection effort and costs.
Number of "collections"	2 Ex: BnF Election 2009	Librarians, collection development: plan and organize specific harvesting projects, design higher level of granularity for collection management, description and access purposes (resource discovery).
Number of "crawls" or "jobs"	7 Ex:Elec2009-20090403000-pre.host2.1	Librarians, collection development, IT Administration, Preservation Planning; define and package collection orders at large scale.
B- Collection processing		
Number of files (URI)	1.5 billion files	Collection characterization and monitoring (Librarians/Technicians), IT Administration, Preservation Planning, Management.
Size of data (GB, TB, PB)	10.2 TB	Collection characterization and monitoring (Librarians/Technicians), IT Administration, Preservation Planning, management.
Number of (W)ARC files	110	IT Administration, Preservation planning.
Number and distribution of MIME types	70% text/html	Collection characterization, IT Administration, Preservation planning.
C- Collection management		
Number of websites which disappeared for a given period and collection (% of loss)	35% of French 2002 Election websites disappeared	Librarians, Management: Evaluate risk of not doing anything and the need to take fast action.
Number of sessions or visits / second (online) or / month (in house access), Number of viewed pages	35 35 981	Librarians, Management: Evaluate usage and usage value
Average cost of a harvested seed: hardware + direct costs + software + labor.	€51	Management, Librarians, IT Administration: Evaluate costs.

Conclusion: Lessons learnt – What comes next?

Towards a definition of heritage value for digital material

With regard to these different elements, we can assume that in the future other means of measuring the importance or the value of digital heritage will have to be invented. For example, it is possible that certain preservation operations will save thousands of documents efficiently and cheaply, without even considering their contents, even though the long term use value of these collections is not at all proven.

Conversely, a small quantity of digital items could be identified as being of high use value and hence the object of much costlier preservation. Thus, we would create a sort of store-room of digital documents, intended for the preservation of rare or precious items, or items of a particular interest, and whose more expensive treatment would be accepted.

We must underline the fact that these decisions will be taken in terms of risk assesment, and that in order to manage risk correctly precise metrics will be necessary. These guidelines should not only concern evaluating the collection itself, but also the preservation actions undertaken; what is the loss rate associated with these actions? How do we evaluate, beyond the loss of data, the loss of functionalities or of particularly significant information? Will it be enough to conserve the contents of these web archives, or should we equally evaluate the quality of end-user experience, for example the navigators used to search the web at a given moment in time?

From a digital heritage conservation point of view, these questions are representative of the stage we find ourselves at today. The work conducted at an international level, and notably by the IIPC consortium for web archiving (and around the standardization of OAIS for digital items in general) has enabled us to gain a deeper understanding of the issues connected with harvesting and ingest. Today we are exploring the question of access and also that of enhancing special sets of data lost in mass or large scale collections. It will nevertheless remain difficult to have a precise understanding of the problems of digital preservation until we have had to face the next step; that is, effective migration or emulation campaigns. Observation of past experience has often led us to take an extremist approach to the problems of preservation. The harvesting of information (metadata, indicators) has been perceived as a reassuring factor - making things easy, enabling us to guarantee that we will have all the technical means of preservation at our disposal. But for the moment we do not know if, on the one hand, we really need such a level of detail and, on the other, if it is economically feasible.

The definition of these metrics presents an essential dimension in the setting up of a digital preservation activity, because they form the sole basis on which we can understand a digital heritage collection. As we gain a better understanding of the collection, we can equally develop preservation strategies, objectives and methods. Guidelines for digital preservation metrics will form a foundation of reference upon which we can construct the interaction between the players, the maintenance, the evolution of the preservation system and, ultimately, a consistent definition of the digital collection.