HAL
open science

Preservation Is Knowledge: A community-driven preservation approach

Sophie Derrot, Louise Fauduet, Clément Oury, Sébastien Peyrard

To cite this version:

Sophie Derrot, Louise Fauduet, Clément Oury, Sébastien Peyrard. Preservation Is Knowledge: A community-driven preservation approach. 9th International Conference on Preservation of Digital Objects (iPRES), Oct 2012, Canada. p. 1-8. hal-00925315

HAL Id: hal-00925315

https://bnf.hal.science/hal-00925315v1

Submitted on 7 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Preservation Is Knowledge: A community-driven preservation approach

Sophie Derrot
Department of Legal Deposit
sophie.derrot@bnf.fr

Louise Fauduet
Department of Preservation and Conservation
louise.fauduet@bnf.fr

Clément Oury
Department of Legal Deposit
clement.oury@bnf.fr

Sébastien Peyrard
Department of Bibliographic and Digital Information
sebastien.peyrard@bnf.fr

Bibliothèque nationale de France (BnF, National Library of France)
Quai François Mauriac
75706 Paris Cedex 13

ABSTRACT
In the beginning, SPAR, the National Library of France's repository, was designed as the OAIS softwarified. It was intended to be a "full OAIS", covering all preservation needs in one tidy system. Then as its potential revealed itself across the library, high hopes arose for a do-it-all digital curation tool. Yet in day to day preservation activities of the BnF, it turns out that SPAR's growth takes a practical approach to the essentials of preservation and the specific needs of communities. Renewed dialogue with producers and users has led to the addition of functions the digital preservation team would not have thought of. This is very clear in what has been created to ingest the BnF's web archives into SPAR, giving the community more information on their data, and in what is taking shape to deal with the BnF's administrative archives, adding new functionalities to the system.The difference between what preservations tools and what curation tools should be at the BnF will have to be examined over time, to ensure all the communities' needs are met while SPAR remains viable.

Keywords
Digital Curation; Preservation Repository; Web Legal Deposit; Digital Archives.

1. INTRODUCTION: BUILDING A REPOSITORY
In the beginning SPAR was designed as a comprehensive digital preservation tool. But we had to reduce its initial scope, and ended up using it for wider purposes than preservation.

1.1 The Original Vision
The National Library of France has been working on building a digital repository to preserve its assets since 2005. This project, called SPAR (Scalable Archiving and Preservation Repository), is intended to be as comprehensive a digital preservation tool as possible. Quite logically, it initially encompassed all the various aspects of digital preservation:

- **Full range of functions.** SPAR meant to implement all the OAIS entities that could be automated: ingest workflow through Ingest, Storage and Data Management functions; dissemination workflow through Storage, Data Management and Access functions; last but not least, a preservation workflow through Preservation Planning and Administration interfaced with the aforementioned workflows.
- **Full range of assets.** SPAR aimed at storing and preserving a very wide range of assets with heterogeneous legal statuses and technical characteristics, from digitized text, image, video and audio content to digital legal deposit, digital archival records and databases, and third-party archived content.
- **The range of preservation levels.** On this double workflow- and content-oriented approach, SPAR aimed at allowing all possible preservation strategies (bit level refreshment and media migration, format migration and emulation) depending on the legal and technical aspects of the corresponding asset.

1.2 Making It Feasible: Prioritizing the Developments and Tightening Up the Scope
This long-term vision could not be achieved in a fully-fledged system and organization in a single run, so the problem and vision had to be split into discrete, manageable, prioritizable bits. This resulted in two aspects:

1.2.1 Splitting the Functions: a Modular Approach
SPAR was designed as a set of interrelated modules, which allowed the system to be developed and updated on a per-module basis. Each OAIS entity was fully implemented as an autonomous module in the system, which communicates with other modules through standard RESTful web services. But all functions did not have the same urgency: before assessing any preservation plans on objects, they first had to be ingested in, and accessed from, a repository. Thus, the development of the Preservation Planning module had to be delayed.

1.2.2 Segmenting the Document Sets: the Tracks and Channels
The preservation policies differed depending on the documents:

- **Legal aspects**: the digital assets to be preserved can be subject to various legal frameworks: legal deposit law; archival records preservation and curation duty law; intellectual property laws and their exceptions for heritage institutions; convention with third party organizations for third party archiving; donations; and so on. Depending on the legal framework of the assets, the library will not be allowed the same range of actions to preserve them.
- **Life cycle management issues**: sometimes it is crucial to have the ability to fully delete all the versions of an AIP in a repository for legal purposes (e.g. for archival records); sometimes it is the exact opposite, with a guarantee that no

deletion of any "version 0" will ever be done (e.g. for born-digital legal deposit); finally, in some cases this might change over time (e.g. digitization, depending on the condition, rarity and complexity of the source physical document);

– **Preservation strategy / Significant properties**: sometimes the content and layout must be preserved (e.g. digitized books), sometimes the top-level priority is the intellectual content (e.g. some archival records), sometimes the user experience is almost as important as the content itself (e.g. "active content" like video games, or born-digital heritage like web archives).

These assets could be grouped in different ways, but few were really satisfactory. Grouping them **by document category** was not very efficient, because different policies could be applied to the same kind of document depending on what is the National Library of France's obligation to preserve it. For example, a born-digital asset will not necessarily be managed the same way if it has been ingested as Legal Deposit or submitted by a third party organization. Grouping the assets on the basis of the **curation services** responsible for them was deemed incompatible with long-term preservation as it would be based on the organization chart, which frequently changes over time. Finally, **a legal framework distinction** seemed well-suited but insufficient, since the same legal framework can be applied to objects with heterogeneous technical characteristics.

However, all these aspects were to be taken into consideration somehow. In other terms, the problem was to find the right balance between the legal, technical and organizational aspects.

This was achieved by grouping the assets into **tracks and channels**. Each track had a set of digital objects belonging to the same legal framework and overall curatorial characteristics, and targeted at a particular user community. Example of tracks included:

– Preservation of digitized books, periodicals and still images
– Audiovisual content
– Web legal deposit
– Negotiated legal deposit
– Archival records preservation
– Donations and acquisitions against payment

Each track is then subdivided into one or more channels, which group together assets with homogeneous technical characteristics.

The first track and channel to be developed was the digitization of books, periodicals and still images, for pragmatic reasons: achieving a critical mass of archived objects very quickly to secure preservation budgets; and achieving a good proportion of the metadata management needs by coping with the best known – and thus most documented – content.

1.3 Making It Real: Back to the Reality Principle

When developing the core functions of SPAR, the team quickly faced huge delays in developments, partly because of the "research and development" aspect of the project and the very specific needs of the BnF in terms of scale, performance and variety of data objects. The functional scope had thus to be reduced. This choice was made on the basis of two criteria:

– Where were the development challenges and failure risks highest?
– What could be abandoned, at least for the moment, while maintaining an up-and-running consistent workflow?

The Access functions were therefore abandoned, as both the most risky part and the dispensable one. For the digitization preservation track alone, the BnF's needs in terms of AIP to DIP transformations (thumbnails, low and medium resolution for web browsing, PDF downloadable content, etc.) were very hard to scale up to the mass of collections at stake (1,5 million DIPs).

From the perspective of our aforementioned different repository workflows, the Ingest, Storage and Data Management modules had priority over the Access and Rights management ones. The library Information System already had existing, though perfectible, applications to manage the digital library and the rights management part. So the scope of our Access module was reduced to the mere dissemination of AIPs. The access and rights management functions were reported to the Access existing applications and Designated User communities for each track.

1.4 It's Alive! Making It Run and Keeping It Growing

With the aforementioned phasing methodology and scope reduction, SPAR went operational in May 2010 for its first core functions and track. From then on, the developments strongly focused on ingesting new content by working on new tracks and channels:

– **Third party storage** (summer 2010): functions to receive content from outside the library
– **Audiovisual track**: audio and video digitization, and CD-audio extraction (spring 2011): audio and video files analysis functions, and management of complex structures such as multimedia periodicals;
– **Web legal deposit** (spring 2012): management of container file analysis (especially ARC files; see below)

Advanced systems administration functions were also added during the first year, and they mostly consisted in helping the IT team manage workflows as efficiently as possible, e.g. to plan mass AIP dissemination and mass fixity checks.

In other terms, the development policy was centered around SPAR as digital library stacks: optimizing the ingest workflows, receiving new kinds of assets (and developing the functions required to do this). This resulted in an increased shared knowledge between curators and preservationists. For each new track, during the design stages, this was initiated with the exchange of knowledge about the digital preservation tool on one hand and the assets at stake and user community needs on the other hand. However, this knowledge of the preserved assets was unexpectedly increased by the preservation tool itself in action.

1.5 Using It: a Digital Collection Knowledge Utility?

The first concrete effect SPAR had on collection curation was indeed the increased available knowledge that was gained on the ingested digital assets, especially regarding their history and overall technical characteristics. The audiovisual track was a good example of such added knowledge, acquired during the tests:

– **Image compression problems:** the curators discovered that some CD boxes and phonogram image shots were LZW-compressed, a format considered risky at the BnF because there was no in-house expertise on it. These images had to be de-compressed before they could be ingested.

– **Unexpected video frame rate structure:** unorthodox 15 frames-GOPs (Group of Pictures)[1] and even variable ones were found. As the content could all the same be displayed, it was decided to ingest and preserve them "as is" but keep all these characteristics in the repository metadata where they could be tracked down.

These two facts were unknown to the library's audiovisual content curators, since they had no impact on the rendering. In this way SPAR's file analysis functions[2] allowed increased knowledge of the collection's technical characteristics. From a long-term perspective, it lowered preservation risks by removing some risky features (e.g. compression) or documenting them (e.g. the GOP) so that the corresponding files could be specifically retrieved in the future.

These features were made possible by SPAR's data management module, which documents nearly all the information required for our AIPs (technical characteristics and file formats, operations performed from creation to the present, policies for ingest and preservation, structure and basic description of the intellectual content) in the form of a RDF database accessible through a SPARQL endpoint [5].

In the end, the design and testing was a very special moment where curators found SPAR gave them a better grasp of the nature and arrangement of their collections. This demonstrated one particular benefit of SPAR where the primary aim was not preservation but rather knowledge of the assets, and therefore curation. This aspect gained even more momentum in the web archives track and the digital archives track.

2. WEB ARCHIVES

2.1 A Track with Very Specific Needs

Since 2006, thanks to an extension of its mission of legal deposit, BnF is mandated to collect and preserve the French publications online [6]. The whole set of data publicly available on the French Internet is concerned: videos, public accounts on social networks, blogs, institutional websites, scientific publications, and so on. BnF uses robots (crawlers) that harvest data from the web and store it in ARC files[3]. The major characteristics that guided the development of the web archives track in SPAR were determined by the specific legal and technical status of these collections:

- legally: long-term preservation, forbidding the deletion of the data, the obligation of preserving the original documents as collected and, at the same time, to give access to the data ;
- technically: data which result from an automatic crawl and even from a succession of different production workflows (by the BnF but also by others partners, by different crawlers, etc.), a wide range of formats and objects.

Of course, the digital legal deposit track's design benefited from the development and reflections on the pre-existing tracks (audiovisual and digitization tracks), and will in turn nourish the next ones (third-party, negotiated legal deposit and administrative tracks). For example, as opposed to the previous tracks, the legal deposit one was bound to strictly forbid the modification or deletion of the original data objects: what the BnF collects by legal deposit must be kept and preserved for access. This question also concerns the administrative archive (see below).

Another example is the preservation of the user experience. For the web archive, not only the content itself, but also its environment of consultation matters; this is not the case for the digitization preservation track for books, periodicals and still images, where content is predominant. To this end, the crawler declares itself as a browser; in order to ensure the harvesting of the content as it was offered to the user. The access to the archive is by an embedded browser and the data must be collected and preserved to enable it to be displayed as on the live web.

2.2 The Challenge of Diversity

It is planned for the web archives to enter SPAR in the automatic legal deposit track. In a way, this track is probably the one which is the most deeply linked with the basic aims of SPAR. The obligation of long-term preservation is impossible under the current conditions of storage of the collections (hard drives and storage bays with no preservation system), and SPAR is the only way for the Library to fully perform its duty. In addition, the diversity of these collections increases the difficulty of preserving and knowing them; only a system dedicated to the treatment of digital collections could permit us to curate such objects.

During the implementation of this track, solutions to several technical challenges had to be found. One of the main issues for web archives preservation is the lack of information on harvested file formats: the only available one is the MIME type sent by the server, which is frequently wrong [7]. To this end, the developments included the design of a Jhove2 module for the ARC format[4]. It is able to identify and characterize ARC files but also the format of the files contained within them. This tool will bring the librarians unprecedented knowledge on their collections. Along the same lines the "containerMD" metadata scheme[5] was implemented to allow the recording of technical information for container files.

BnF web archive collections are made of several data sets which came from different harvesting workflows [8], in different institutions with various practices (the BnF, the Internet Archive foundation, Alexa Internet which worked with IA). SPAR was a natural choice for preserving these web archives, but some adjustments were necessary on both sides, and particularly the homogenization of the different collections into one data model. Inside the track, five channels were distinguished, according to the workflow using for the harvest. Not every channel has the same level of description and metadata. The librarians knew from the beginning the major differences between the channels, but this knowledge was markedly improved by the implementation of the track and the necessary work of homogenization.

---

[1] The Group of Pictures is a way to document how the moving image stream is divided into full frames and, if any, intermediary frames that only list the differences from the next frame in a predictive fashion. See http://en.wikipedia.org/wiki/Group_of_pictures.

[2] SPAR identifies formats with a Java-packaged File UNIX command, and analyses image and text with JHOVE, audio and video with Mediainfo, and ARC container files with JHOVE2.

[3] ARC is a container format designed for web archives (see http://archive.org/web/researcher/ArcFileFormat.php). Its evolution, the WARC format, is an ISO standard (28500:2009)

[4] See https://bitbucket.org/jhove2/main/wiki/Home. Development of a WARC module for Jhove2 is currently performed by the Danish Netarchive.dk team.

[5] On containerMD, see http://bibnum.bnf.fr/containerMD.

2.3 Knowing Collections by Implementation

The SPAR team is now close to the end of the implementation of the digital legal deposit track, which began two years ago. This provides an opportunity to consider the choices made at the beginning of this work.

RDF was chosen as the indexation model in SPAR. The triple-store capacity is limited, and the stand was taken not to index some data of the ARC files, especially the associated files. During a crawl performed by Heritrix and NAS, files are produced with reports and metadata about the crawl (crawl log, hosts reports, seed list); the large size of these files made their complete indexation impossible. Thus it is impossible to obtain by a SPARQL query the list of the harvest instances containing a certain domain name. This was a conscious choice made during the development of the track, and therefore a known limit of the knowledge about the collections.

On the other hand, a lot of metadata are indexed and therefore can support a SPARQL query. Especially, SPAR ingests reference information about agents performing preservation operations, which can be performed by humans (administrators, preservation experts), software tools (identification, characterization and validation tools) and processes in SPAR (such as the ingest and package update process). Performing these requests allows precious statistic, technical or documentary information to be retrieved about the collections:

- for example, the list of the crawlers ("agent") and the version used by channel can be produced by querying the agent linked to the harvest event with a role of "performer":

Table 1. Response to a SPARQL query on crawling software tools for each channel

channelId	agentName
fil_dl_auto_cac	Heritrix 1.10.1
fil_dl_auto_cac	Heritrix 1.12.1
fil_dl_auto_cac	Heritrix 1.14.0
fil_dl_auto_cac	Heritrix 1.14.2
fil_dl_auto_cia	Heritrix 1.14.1
fil_dl_auto_cia	Internet Archive
fil_dl_auto_his	Alexa Internet
fil_dl_auto_htt	HTTrack 3.10
fil_dl_auto_htt	Alexa Internet
fil_dl_auto_htt	HTTrack 3.30
fil_dl_auto_nas	Heritrix 1.14.3
fil_dl_auto_nas	Heritrix 1.14.4

- another example is the list of harvest instances with "elections" in their title or description:

Table 2. Response to a SPARQL query on harvest instances concerned by the electoral crawls

Harvest definition	Title
ark:/12148/bc6p03x7j.version0.release0	BnF elections 2002
ark:/12148/bc6p03z7s.version0.release0	BnF elections 2004
ark:/12148/bc6p03zd5.version0.release0	BnF elections 2007

At the end of the implementation process, testing the possibilities of SPARQL queries on this track allowed the discovery of a few bugs or mistakes. But most of all, it gave the opportunity to fully consider the tool offered for the management of the collections.

The heterogeneity of data models between web archives from different periods was a strong obstacle that prevented from having a common view on the BnF collections. The alignment of those data models and the possibility of requesting all collections the same way thanks to the data management module will permit getting similar metrics for all kind of assets. In that way SPAR will help providing the BnF the statistics and quality indicators necessary to measure and evaluate its collection. A list of these indicators is currently designed by a dedicated ISO working group, whose draft recommendations influenced the implementation of the web archives track[6].

Testing the preingest phase for the test dataset also allowed the application of comprehensiveness tests. Each ARC metadata AIP contains a list of all ARC files produced by the harvest instance, as the outcome of a harvest event. Automatically comparing such lists with the ARC data files actually ingested in SPAR may prove very useful with old collections, for which there is a risk of losing data. It ensures too that incomplete or defective datasets cannot enter SPAR, which could otherwise be problematic for the preservation process. This new feature has been added to the administration module GUI.

2.4 Outside of SPAR

SPAR is the natural way to preserve the web archives over the long term. But in the meantime, several migration and packaging operations are performed outside of SPAR, which could have been thought of as typical preservation operations. For example, the BnF is planning to migrate all its ARC files to WARC files, thanks to specific migration tools. These tools will not be part of the SPAR workflow, but will be external. However, all the operations on the collections will be documented in the system, as the PREMIS data model, the cornerstone for SPAR's RDF data model, allows the monitoring of each "Event" related to a file or a file group. The traceability of this kind of operation is key information to the curation of digital collections.

On the later crawls, the data harvested by the Heritrix are prepackaged and enriched by metadata on the harvest by the curator tool, NAS. So the majority of the metadata on the harvest itself is pre-existing and therefore quite easily controlled by the librarians. This could be seen as easier on a daily basis, but it is also restrictive because every modification of the external tool must be made in the perspective of the ingest in SPAR. It forces the librarians to consider their collections from a preservation point of view and reinforce the consistency of the collection.

3. A DIFFERENT KIND OF COMMUNITY: ARCHIVES IN THE LIBRARY

3.1 Yet Another Track

During 2012, the SPAR team has been focusing on the ingestion of archives. The plan is to build experience with the BnF's own documents, with a view to expanding its third-party preservation offer in the process, to records and archives in other institutions. In preparing the requirements for a new tender to further develop the system, starting this fall, the preservation team is learning yet again how taking into account new producers and designated communities is pushing the services of the Archive, and even its philosophy, in new directions.

[6] The ISO TC46/SC8/WG9 is currently working on a Technical Report (ISO TR 14873) on Statistics and Quality Issues for Web Archiving that will be validated and published within a year. See also [2] on the question of web archive metrics.

3.1.1 Different Legal Requirements

Although France has promulgated a unified code of law for its cultural heritage, the *Code du Patrimoine*[7], in 2004, it does not imply that a library could pick up archives and know what to do with them. And yet, the BnF has been producing records of its activities, and has been managing its own administrative archives, from the paper ages to the digital times. It has created a dedicated bureau to do so, recruiting archivists trained in the specificities of records management and the curation of historical archives, regardless of their medium.

Thus, in order to preserve the growing digital part of these archives, the SPAR team is now dealing with a new kind of producer and user community, and information managed under different rules of law. In the system, this translates into the creation of a new "track" for "administrative and technical production".

The main constraints that differ widely from the digital preservation team's previous endeavors with digitization and legal deposit stem from the added complexity of the information lifecycle: there is a much higher chance that information may be accessed and reused to create new versions of documents, and, above all, it may, and sometimes must, be deleted. The law on public archives requires that, once they are no longer in active use, documents that are not important for administrative, scientific, statistical or historical purposes should be weeded out of archives. Should different service levels then be applied to different stages in the lifecycle? Up to which point can sorting and eliminating records be automated? The role of SPAR in this process is beginning to take form.

3.1.2 A Specific Technical Environment

While acclimating to this different legal context, the digital preservation team also has to take into account an increased variety of documents and data, and specific work environments. The BnF's archives encompass the usual office documents — word processing, spreadsheets, slides and PDFs, — as well as a long trail of varied file formats, and the number of documents not in a format from the Microsoft Office suite increases steadily over the years. The library also produces highly specific records of its activities using specialized business software, such as financial databases or architectural plans.

From the first overview of this "track" in SPAR, it had thus been posited that several separate "channels" would be required to deal with the various types of records from the library's activities, and interact with their different production environments. A choice was made to focus this year on what is supposed to be the most standard of those channels, the one for regular office work records.

Yet there are challenges, given that the documents are stored and classified using proprietary software, IBM Lotus Notes. In addition, the BnF's agents tend to use this software in an idiosyncratic manner, in spite of the library archivists' efforts over the past years to fit it closely to the library's records production. Moreover, it would seem that the designated community for this part of the Archive is the largest SPAR has ever had to serve so far: producers and users of the administrative records are the library agents as a whole.

Their representatives in the working group piloting SPAR's developments are brand new to the process, and bring a new and highly technical knowledge to the building of the repository: the BnF's two archivists have experience in records management and archival law, its Lotus administrator understands the workings of data and metadata in the document-oriented databases. Following the needs of the designated community for this new "channel for administrative production" is again changing the original contour of the SPAR system.

3.1.3 A New Set of Challenges

With the first tender for the development of SPAR's software ending in January 2012, it was decided that a first study of the requirements for the Administrative Channel would serve as an evaluation tool for potential new contractors. In the few months of this first investigation of the needs for the preservation of the BnF's administrative archives, issues emerged regarding which data and metadata to collect, create, preserve and disseminate. For instance, SPAR's team had never had to deal before with

- a greater attention to the issue of integrity and authenticity: the records and archives world is much more concerned with the possibility that a document may be required in a judicial context, where it will be necessary to prove that it has not been tampered with. What this means in a digital environment has yet to be clarified by jurisprudence;
- a lifecycle that may require documents to be accessed and modified in their original production environment, and, later on, in an updated or different business management environment that would have to interpret the original data and metadata correctly, and allow complex use of it;
- a more pressing need for a mechanism to delete AIPs and trace those deletions.

Other institutions and companies have had to solve such problems before[8], but in the context of a library, and at this point in the development of SPAR, they are likely to be the source of a whole crop of new features in the system.

3.2 How to Manage: Verify, Migrate, Delete?

Given that preserving records is not necessarily new business, the BnF did not set out to reinvent the wheel, but existing solutions for records management and digital archiving did not fit the library's preservation plan:

- the core functions of SPAR have been designed to be generic, i.e. deal with information packages from all tracks and channels with the same processes. Introducing a whole new system was not considered an option;
- the requirements for the modernization of the French administration have first focused on a specific set of records that do not match the diversity of data in the BnF's Lotus Notes bases, nor its specific structure.

There is a national standard for the exchange of archival data ("Standard d'échange de données pour l'archivage", SEDA[9]) that the BnF will implement to deal with the messages and metadata attached to information transfer between producers, Archive and

[7] The latest version of which is available, in French, at http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006074236 (accessed 21 May 2012).

[8] Regarding rendering office documents for instance, Archives New Zealand's recent report is illuminating [4].

[9] Schemas, tools and profiles are available, in French, at http://www.archivesdefrance.culture.gouv.fr/seda/ (accessed 14 May 2012). A version 1.0 of the standard is in the works.

users. However, to create an interface between Lotus Notes and SPAR, this standard might not be fitting or necessary.

Moreover, the integrity of the BnF's Lotus databases is secured by multiple replications. The role of SPAR in the preservation of administrative production was rapidly defined by the working group as long term preservation of archives, not bit level preservation in the records management processes. Which of the records management processes, then, have to be maintained when the lifecycle of the records brings them to the point when they are ingested into the SPAR repository?

3.2.1 The Problem with Signatures

The BnF's archivists and IT specialists have secured authenticity in the library's records management through user authentication, digital signatures — to prove a record's origin, and access control lists — to manage access rights to the application, document, view and item levels. Whether this information can, and should, be carried over to the SPAR repository is a question the BnF has to research further. At this point in the specifications of the future Administrative Channel, it seems that it would be a Sisyphean task to renew the certificates associated with the signatures regularly since the certificates have a lifetime of a few years, and most of the BnF's archives reaching SPAR are to be preserved indefinitely.

It may however be useful to verify each document's signature at the moment the documents are transferred from the Lotus databases to the first stages of the ingest process. The signature files themselves might even be included in the METS manifest of the information packages if their durability can be proved. It seems likely, however, that the main assurance of the records' authenticity will come from sustaining and demonstrating the trustworthiness of SPAR's processes. This actually agrees with the practices of the producers and users of this Administrative Channel: the BnF's archivists rely as much on available documentation as on their skills in analyzing records for clues about their provenance, authenticity and integrity. In the working group, they told the preservation team they did not expect digital records to conform to an authenticity standard that has never been required in the paper world.

3.2.2 Conciliating Preservation and Access: Instant Migration

As can be expected in a large institution such as the BnF, constraints about number of users and budget, licensing fees in particular, make it difficult to switch to the latest and most easily preserved technologies. The library still relies on the 2003 Microsoft Office Suite, for example, with only binary formats available so far. Furthermore, the diversity of the library's activities means that no limit can be imposed on the file formats used, although the use of Word, Excel and PowerPoint files as attachments is facilitated, and represents about half of the files present in the databases.

The Administrative Channel processes must guarantee that the archived documents can be rendered again at any time in the Lotus Notes interface, in all their diversity. Which means that the specific structure of the Lotus document-oriented databases must be preserved as well: each document is stored in a series of fields, regardless of what could be considered data, or metadata. The items in a document encompass detailed provenance information, as well as rich content and attachments. Lotus provides an export and import function in a proprietary XML format, DXL, that may solve the issue.

Meanwhile, the service level for these documents in SPAR must be better than the bit-level preservation in an extraction in a proprietary XML format, and it must guarantee not only future rendering, but also modification of the data: relying on emulation alone might not be enough. The SPAR team is investigating the following approaches so far (see Figure 1):

- recording the visual aspect of the original document in a standardized format, using Lotus' PDF export capabilities for instance;
- taking the encapsulated files out of the DXL export of the document, making them easier to identify, characterize or migrate over time;
- transforming the remaining data in the DXL files to an open format, such as XHTML;
- making it all apparent in the "USE" attribute of the corresponding file groups in the METS manifest of the information packages.

Historically, files that are considered the focus of preservation are in the file group that has a USE "master". Here, it would correspond to a standardized representation of the Lotus document and the formerly encapsulated files. The Lotus document without its attachments, where all the descriptive and provenance information would remain, would, in its transformed version, make up a file group with the USE "documentation", which designates in SPAR the set of files containing metadata that cannot be entirely incorporated to the METS manifest but should be accessed for preservation planning. This document in its proprietary DXL format would be part of a new type of file group in SPAR, with the USE attribute "original": working with the designated community of the Administrative Channel has made the SPAR team realize that it lacked a USE in its nomenclature for files that are not the primary object of preservation but must be stored for reuse in their original environment.

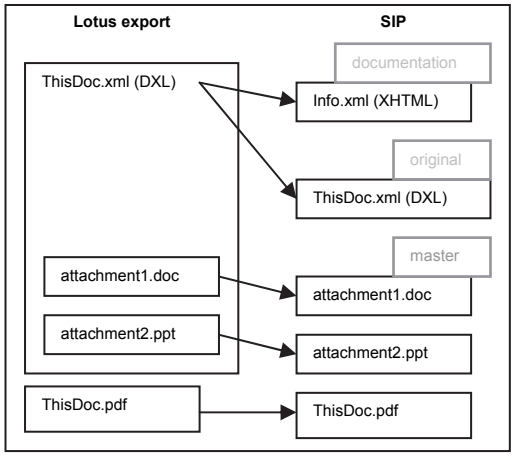

Figure 1. Creating a SIP from a Lotus Notes document

Using a similar logic, it appeared that in order to maintain usability of the Lotus documents in their original environment and to secure a higher service in the preservation process, attached files in proprietary formats could be transformed as well. This would be better accomplished not at the SIP creation stage, which deals with the way the Lotus export is recomposed, but within the system, according to the preservation planning capacities of SPAR at the time of the ingest. For example, a Microsoft Word binary file could be transformed into an Open Document file. The original Word file would be preserved in the information package for dissemination via the Lotus Notes interface, but would be

moved to the file group with the USE "original", while the new Open Document file would now be part of the file group with the USE "master", as the option chosen for long-term preservation actions (see Figure 2).

As for the DIPs, depending on the time and context of dissemination, they could combine files from file groups of different uses. This is yet another function that the SPAR team has had to take into account rapidly as a result of the dialogue with the representatives of producers and users in the Administrative Channel, since the repository so far can only disseminate DIPs that are an exact copy of the AIPs.

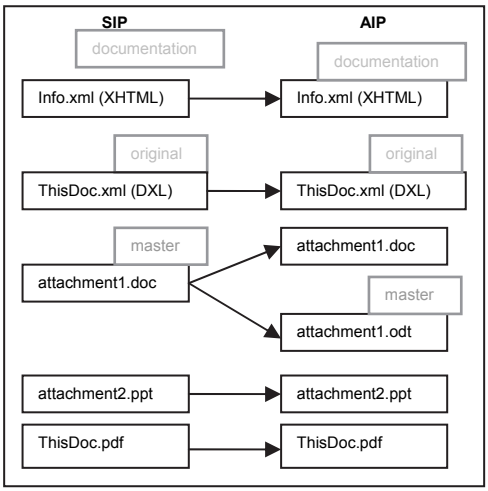

Figure 2. Migrating and moving files from SIP to AIP

3.2.3 Ending the Lifecycle: How to Delete

More flexibility at the access stage was something planned at the design stages of SPAR, that was scaled back because the communities for the first channels had no use for it, and moved forward again when producers and users made the case for its importance in their collection curation processes. Another example of these shifting priorities to serve the community is the deletion function. In the beginnings of the SPAR project, a lifecycle was devised for every AIP in the system: their first version, or version 0, would be preserved forever, as well as the latest one, and the one before, to allow for rollback. The implementation of this model was delayed, all the more since the first channels in SPAR contained collections whose forms were stable and was preservation was infinite.

Working with the records managers and their IT counterparts has shown the SPAR team that the deletion mechanisms have to be much more supple, while remaining simple, because of the high degree of human expert intervention in the lifecycle decisions. Although the documents in Lotus contain information regarding the duration of preservation required that is automatically assigned according to the document type, it cannot be used to pilot lifecycle decisions in SPAR: the intervention of an archivist to decide which documents are part of a closed case and are ready to be archived in the repository is necessary. Similarly, the BnF's archivists must validate all deletions. Moreover, these deletions have to be properly documented.

Given the design of SPAR, a solution might be to submit SIPs describing a "deletion request" event in their METS manifests. This would update the AIPs to include a "deletion processed" event documenting the action in their manifests while ridding them of their data objects, and set off the deletion of all previous versions of the AIPs. In any case, integrating such new and crucial abilities into a functioning system will be an interesting challenge for the end of the year.

4. CONCLUSION: CURRENT ACHIEVEMENTS AND NEXT STEPS

4.1 Coverage of the OAIS Model

In its original conception, SPAR was intended to implement, as strictly as possible, of the OAIS model – indeed both OAIS models, the information and the functional models. Considering what has been achieved, to what extent has this objective been reached?

4.1.1 Information Model

The repository uses the full typology of information in the OAIS information model – but its precise nature, the way it is organized and the level at which it can be found highly differs from one track to another. In the digitization and audiovisual tracks, most metadata are recorded in the METS manifests. These METS files directly express structural metadata, and thanks to other metadata schemes embedded in METS, contain representation information (in MIX for images, textMD for text and MPEG-7 for audiovisual content), provenance and context information (in PREMIS), and descriptive information (mainly in Dublin Core). Fixity (checksums) and reference information (ISBN for books, persistent identifiers for all kind of documents, etc.) are included as well.

On the contrary, in the web legal deposit track, some representation information (MIME types of each contained file) is directly available in the ARC files, but is not described in METS. Moreover, METS files contain very few structural metadata, as the structure of web archives is already recorded in the hyperlinks present in the archived web pages. Descriptive information is only available at a very high level. In the end, it is perhaps in the use of PREMIS for context and provenance that the different tracks are the most similar.

As for rights metadata, which were not identified as such in the first version of the OAIS, they are not described yet in the metadata profiles. However, any descriptive, context or provenance information may be the basis for rights metadata, as they may help deduce the legal statuses of the documents. In fact, the very definition of each track depends on the legal status of the documents in it.

4.1.2 Functional Model

As to the functional model, one might consider that all functional entities have been implemented in SPAR modules – but at very different levels of completion. Modules highly related to collection knowledge and collection storage reached a high level of achievement: the ingest module extracts and computes a large number of metadata, which can be requested by the data management module. The storage and “storage abstraction services” modules are able to choose dynamically between different media storage and on what physical sites data should be stored. On the other hand, the access entity functional scope has been reduced to the bare minimum: to extract requested AIPs as they are from the system.

Yet the SPAR system has never been thought as a dark archive or a black box, but as an accessible system. However, designing a generic access module, able to create custom DIPs for digitized books, video games as well as web archives, is an objective currently beyond reach – and too ambitious for a project which was intended to show concrete results in a few years.

Finally, there is still work to be done on the administration and the preservation planning sides. New administration features are added each time new tracks and channels are developed, but a lot of improvements can be made on interfaces and ergonomics. These enhancements will probably be accelerated by the growing number of users as new challenges appear.

The preservation planning aspect is also less developed than what is expected in the OAIS model. On one hand, many functionalities of SPAR help design preservation strategies. Knowledge gathered at ingest, especially during identification and characterization processes, represents the cornerstone of a preservation strategy. On the other hand, we still do not have any tool to match automatically formats to preservation strategies. One of the next steps would be to let the system interact with format repositories like UDFR.

4.2 Next Steps

The second main phase of development will therefore extend the scope of SPAR in several directions:

- ingesting new types of collections. The administrative archives track is the next one to be integrated; electronic periodicals acquired by the BnF, e-books and other digital-born documents collected through legal deposit will have to follow.

- improving existing tracks, by adding new channels for instance. These new channels could be based, not only on the legal and technical statuses of the documents, but also on their scientific, heritage or financial value – taking into account the fact that this value may evolve through times.

- opening the repository storage and preservation facilities to the BnF's national partners using SPAR's third-party archiving track – in the heritage realm or not. This is probably less a technical than an organizational issue: to whom should these services be offered? At what cost? Who will be liable in case of problems?

- defining the professional profiles involved in the development and the daily use of SPAR. Until now, the development of the SPAR project has been followed on a day-to-day basis by two kind of professional profiles: IT engineers (developers and analysts) and "digital preservation experts", i.e. librarians with a strong technical knowledge, who are in charge of assessing and maintaining metadata and data formats. Representatives of the Producers and User communities are also involved in the design stages of their tracks. However, a larger permanent working team is needed to maintain the live system while the developments continue. The content curators need to be more involved in the preservation of the collections they helped creating. Otherwise, digital collection curation and preservation will never be considered mainstream librarian activities.

The human part of digital preservation has probably been the least studied up to now, even though a working group called ORHION (Organization and Human Resources under Digital Influence) has been since 2009 dedicated to these issues [1 and 3]. A whole librarianship activity needs to be built around the SPAR system. Who will manage the system? Who will be able to send requests to the data management module? Who will be able to update metadata? Who will decide on preservation actions? This points to a general problem about the Designated communities and the frontier in their daily work between preservation and curation activities: is SPAR designed to be a digital curation tool as well as a preservation repository, or must new tools be developed as new needs are identified?

In its first design, SPAR was supposed to be a fully integrated digital preservation system. It is now a secure storage repository that offers its communities the ability to know and to manage all their digital collections. Some preservation actions happen outside SPAR– but the system is able to document them. On the other hand, SPAR makes a lot of information available for the first time, giving insight and control on the digital collections it holds. From this point of view, SPAR is redesigning the frontiers between preservation systems and curation tools at the BnF, reinventing librarianship for digitized and digital-born collections.

5. REFERENCES

[1] Bermès, E. and Fauduet, L. 2010. The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France. *International Journal of Digital Curation*, 6, 1 (2011), 226-237. [http://www.ijdc.net/index.php/ijdc/article/view/175/244]

[2] Bermès, E. and Illien, G. 2009. Metrics and strategies for web heritage management and preservation. In *Proceedings of the 75th Congress of the International Federation of Library Associations* (Milan, Italy, August 23-27, 2009). [http://www.ifla.org/files/hq/papers/ifla75/92-bermes-en.pdf]

[3] Clatin, M., Fauduet, L. and Oury, C. 2012. Watching the library change, making the library change? An observatory of digital influence on organizations and skills at the Bibliothèque nationale de France. To be published in *Proceedings of the 78th Congress of the International Federation of Library Associations* (Mikkeli, Finland, August 11-17, 2012).

[4] Cochrane, E. 2012. *Rendering Matters - Report on the results of research into digital object rendering*. Technical Report. Archives New Zealand. [http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering]

[5] Fauduet, L. and Peyrard, S. 2010. A data-first preservation strategy: data management in SPAR. In *Proceedings of the 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, September 19-24, 2010). [http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/fauduet-13.pdf]

[6] Illien, G. and Stirling, P. 2011. 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 Congress of the International Federation of Library Associations* (San Juan, Puerto Rico, August 13-18, 2011). [http://conference.ifla.org/past/ifla77/193-stirling-en.pdf]

[7] Oury, C. 2010. Large-scale collections under the magnifying glass: format identification for web. In *Proceedings of the 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, September 19-24, 2010). [http://netpreserve.org/about/Poster_ipres2010_webarchivefileformats_oury.pdf]

[8] Oury, C. and Peyrard, S. 2011. From the World Wide Web to digital library stacks: preserving the French web archives, In *Proceedings of the 8th International Conference on Preservation of Digital Objects* (Singapore, November 1-4, 2011), 231-241. [http://getfile3.posterous.com/getfile/files.posterous.com/temp-2012-01-02/dHqmzjcCGoexvmiBzJDCyhrlhIgswoffzvsfnpEAxjHFEesarvwahEHrmyvj/iPRES2011.proceedings.pdf]