Destination: Shared Repository
The National Library of France's Journey to Third-Party Archiving

Louise Fauduet
Department of Preservation and Conservation
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
louise.fauduet@bnf.fr

Sébastien Peyrard
Department of Bibliographic and Digital Information
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
sebastien.peyrard@bnf.fr

ABSTRACT

The SPAR repository project started as a way to make preservation easier, cheaper, and more effective for the National Library of France (Bibliothèque nationale de France, BnF). At the time (early 2000s), the BnF used different storage media and technologies across the library, and had no unified responsibilities and processes. When the decision was made to overhaul the infrastructure and the software to have one repository to replace them all, the project designers reflected quite naturally that such an effort could and should benefit other libraries that had similar scalability issues.

There were many obstacles to overcome after that generous impulse. First, finding the BnF's niche in the digital preservation landscape, at a time when it was growing and evolving fast — as it still is. Even when focusing on the heritage sector, there are several other repositories or planned systems that have a national vocation, within the archives or higher education communities for instance. Then came the matter of combining the needs of the library itself and the design necessities of third-party archiving to create a repository that would be scalable, trustworthy and open. Last but not least, the BnF is continuously refining the way the repository can best serve its clients. The most accessible function for partners is bit-level preservation, and any extra step toward comprehensive preservation has to be balanced with available resources and tiered prices. As the first clients come in, prices and processes are still in flux, and vulnerable to policy changes.

Keywords

Digital repository. Third-party archiving. Cost models.

1. INTRODUCTION: ARCHIVING AS A LIBRARY?!

1.1 What we talk about when we talk about archiving

The French language is fraught with nuances that do not quite translate into English. One such problematic word is "archive": amongst cultural heritage professional circles, archives are first and foremost the place where records go when they grow up and become "permanent", while the French word "*archive*" is applicable to any stage of the archiving process. And archives are not the libraries' domain, usually, except when authors' papers lose their way and enter library collections. What's more, in France, the term "*archivage*" can be used for records management, archiving and preservation, or third-party archiving where it is called "*tiers archivage*". In other terms, the same "archive" term can be be used for curation or for preservation, depending on the context. The same problem occurs with "tiers archivage", which means third-party archiving but can be applied for preservation as well as archiving services. Last but not least, a term like "Web archiving" is used to define something which, in France, comes under Legal Deposit law.

In other terms, "*archivage*" can be applied to different professions, skills and activities (archives, libraries), and to different legal statuses (administrative production or publication). Moreover, when it comes to third-party preservation of archive records, any institution entitled by the central services for French archives to perform such activities, can paradoxically do archive preservation without being an archive center in the first place[1].

France's deposit and archiving laws may be puzzling as well. Certain types of materials have clear destinations in paper and digital forms, others are not constrained or are unclear. For instance, the publication of a research center in a University can be considered a publication and thus subject to legal deposit, but can also be considered a public archive record as a product of an agent of the state. Which legal system applies to it depends very much on the heritage institution that will take on the responsibility of preservation

1.2 France's digital preservation landscape

If we focus on the cultural heritage sector, France's digital preservation landscape is shared between a few large institutions, especially in the Ministry of Culture (public libraries and archives) and the ministry of Higher Education and Research (universities, research centers and datacenters). The administrative distinction between those two ministries is very relevant in understanding why the three main systems in France involved in heritage digital preservation are the BnF's SPAR system, CINES' PAC platform and IN2P3's datacenter for scientific experimental data. This rather centralized landscape fits the recommendations of the report called "Strategic orientations for digital libraries"[2], produced by the president of the BnF under the auspices of the Ministry of Culture in early 2010, which insisted upon lowering public costs in order to produce economy of scale. On the archives side, the landscape is more parceled out: the French Archives ministerial services have a role of technical recommendation, control and advice, but some local archive centers have developed

[1] This is the case of the CINES and BnF who, among other organizations, received the grant to ensure archive preservation.

[2] Called in French "Schéma Numérique des bibliothèques": http://www.enssib.fr/bibliotheque-numerique/document-48219.

their own solutions. The VITAM[3] project intends to provide a large scale solution for the National Archives and the Archives of the Ministry of Foreign Affairs in a three-year time frame. Another solution called CDC Arkhinéo[4], targeted at third-party legal archiving, has been developed by the French public institution called Deposits and Consignments Fund ("Caisse des dépôts et consignations") with a strong focus on security and legal evidence. The National Center for Scientific Research has a solution for digital humanities, Huma-Num[5]. At local scale, some repository solutions are being developed, e.g. M@rine developed by two department archive centers[6] And several private firms, some with experience in records management and archiving for banks, for instance, are offering their products to public archives.

In compliance with the strategic orientations mentioned above, the DISIC (Interministerial direction for IT Systems) created in 2011, has a similar mandate of rationalizing the public expenses on IT infrastructures. Its focus on digital preservation, however, will only focus on technical aspects, leaving the key organizational challenges outside its perimeter.

Given the history and context of the SPAR project, the BnF services are somewhat different from the most common shared repository models:

1. Projects that started with a national or local mandate and *ad hoc* governance structure, developing and sustaining the repository for its members (National Digital Library of Finland, HathiTrust), with a partnership model;
2. Projects with a national or local mandate to provide a service to a community, where the service provider has no collections of its own in the repository, and the customers are not part of the board (CINES, California Digital Library's Merritt);
3. Software solutions where the vendor fosters a community of users, either as a downloadable software (e.g. SDB, Archivematica…) or as an online facility (Duracloud, Preservica…);
4. Networks of repositories exchanging copies of their information packages (e.g. LOCKSS networks, Chronopolis, TIPR…)[7].

The BnF sells storage and services, but mostly maintains control over the technical roadmap and the repository governance, as SPAR was first developed for its own preservation needs. So far it is closer to an institutional repository model.

2. MAKING A SHAREABLE REPOSITORY

2.1 Looking for scale

The BnF's main strength compared to other heritage institutions is the size of its own collections. SPAR became operational in May, 2010. As of June 2013, the repository hosts around 1 million information packages, representing over 800 Tb, essentially from the library's digitized collections. Many more hundreds of terabytes from the backlog of digitized collections and from Web archives collections are being ingested. The current storage capacity of the system is 1 Pb, with about 16 times more in terms of slots available in the tape library. This may not be very sizable on the international scale, but it is for example much more than the CINES has budgeted so far for the collections its repository stores, at 40 Tb. There is no doubt that the other repositories will grow, but the BnF's SPAR has a head start given the library's own needs, and thus has already achieved a certain economy of scale regarding storage.

What's more, the software itself has been designed to scale up. By making it as modular as possible, the development team hopes to be able to change any given module (ingest, storage, data management, etc.) according to new progress in technology or new requirements in scale. Another strategy has been to add multiple instances of the most used modules, which deal with SIP preparation and ingest.

Above all, the design for SPAR has been based on the concept of "tracks" and "channels", to organize content and make managing heterogeneous collections easier. Tracks are created according to the legal status and entry mode of the collections they enclose: digitized materials, legal deposit, gifts and acquisitions, etc. A track for third-party archiving has been envisioned from the beginning. Channels are sub-divisions of tracks according to technical challenges and refinements in preservation requirements. With each channel, a new set of service level agreements are negotiated, defining the conditions for ingest, preservation and access.

Thus the logic of the system is to have at least one new channel created for each third party submitting assets to the SPAR repository, with its own set of negotiated parameters. Should the nature of the collections entrusted to the BnF by a third party be varied in its technical composition or in the level of care it requires, then more channels should be added. The upside to this is a high adherence to the needs of the partners; the downside is the extra burden on the BnF's staff and resources each time a new channel must be set up and maintained.

2.2 Looking for standards

Making the philosophy and design of a repository compliant with standards is a key condition to its being shareable. Hence the use of the OAIS standard to design the system and the use of the METS and PREMIS standards for its data model and the preservation metadata of each document. These proved to be invaluable since initiatives can be initiated on the international scale that benefit back to the repository. For instance, the BnF will take part in the Preservation Health-Check Pilot[8] in the course of 2013, whose purpose is to give a risk driven evaluation of the METS/PREMIS metadata stored in the SPAR repository. The BnF could not have been part of such an international R&D projet without standard metadata formats. Another great added value those standards provided was genericity, whatever the kind of

[3] http://www.archivesnationales.culture.gouv.fr/chan/chan/english-version-colloque-archiving-2013.html.

[4] http://www.cdcarkhineo.com.

[5] http://www.tge-adonis.fr.

[6] This solution ensures preservation as well as archive-specific curation functions. Cf. http://www.sicem.fr/index.php?option=com_content&view=article&id=167&Itemid=41.

[7] A recent census of existing preservation repository initiatives can be found in Aligning National Approaches to Digital Preservation Conference Edited Volume. http://educopia.org/sites/educopia.org/files/ANADP_Educopia_2012.pdf.

[8] http://www.oclc.org/research/activities/phc.html.

content was; and, in a longer-term perspective, lower the barrier to making the other systems and initiatives mentioned in 1.2 interoperable with the BnF repository solution to allow distributed preservation over the country.

In addition, efforts were made to use open source software whenever possible in SPAR's own code. The principles are the same as with the use of standards: benefiting from community-approved tools, and adding to them whenever possible (the BnF has commissioned an ARC and a GZIP module for JHOVE 2, for instance), while fostering interoperability.

2.3 Looking for certification

Once the BnF decided to open its services to third parties, it was important to prove its trustworthiness to them. To this end, the BnF has been monitoring the certification initiatives that have started ever since the OAIS was first published, including the TRAC and DRAMBORA check-lists. With the birth of the European Framework for Audit and Certification of Digital Repositories, in 2010, the path to certification is now clearer. However, on top of international certification, the BnF is also concerned with French standards and certifications.

It is currently interested in three 3 parallel certification initiatives:

1. The authorization to preserve third-party archive records, required by French Law for an institution or a firm to be entitled to store and preserve public administrative documents;
2. The French AFNOR[9] Z42-013 standard, which evaluates the technical trustworthiness, security and traceability of the preservation system. It has been transformed into ISO-14641-1 at the international level. (ISO 14641-1:2012, Electronic archiving -- Part 1: Specifications concerning the design and the operation of an information system for electronic information preservation[10]). The corresponding French certification was created as NF-461 in early 2013;
3. The Data Seal of Approval[11], which evaluates the OAIS compliance of the repository

The BnF received the first of these in Spring 2013, after a year of discussions with the central services for the French Archives at the Ministry of Culture, who deliver the authorization.

These efforts have revealed two main issues with the BnF's certification efforts, that will be addressed in the coming months with the help of the person in charge of disaster and risk management at the library, and with new software developments:

- the lack of policy statements at the library level – the preservation policies have so far been discussed and implemented with collection managers directly – and of technical documentation in English. Those documents would be essential to getting a Data Seal of Approval, for instance;
- the low level of security required for the preservation of the library's own collections, compared with the authenticity standards expected in dealing with public or private archives, for instance, due to their potential roles in judicial processes.

Thus, while working on its digitization or Web archives collections, or even on the initial phases of its third-party archiving services, the BnF hasn't invested in time-stamps, tamper-proof hashes, and certified signatures. It is working on these aspects now that the third-party services are attracting more interest.

3. GETTING CLIENTS

3.1 Defining services

The SPAR system has been developed in an iterative way: after the core functions were created in 2008-2009, and after a first track was set up for digitized collections, seen as the most urgent preservation need at the time, the team went immediately on to work on the third-party archiving track.

There were managerial motivations to this decision, as the push for unified repositories at the State level was already being felt. There had also been a trend within the library to generate income from its own services.

For the system designers, this represented an opportunity to make it more generic and customizable. The core of the repository is intended to be as generic as possible: all functions dealing with SIPs, AIPs and DIPs must be available to all tracks and channels, whether they use them or not, so that the ingest, data management, storage and access modules are standardized and easier to maintain. However, to deal with the wide variety of objects to be preserved at the library, specialized pre-ingest modules are created for each channel, in order to turn specific information submitted by the producers into normalized SIPs.

The first pre-ingest module created for SPAR dealt with the highly specific requirements of the BnF's history of digitizing its own collections according to the strict rules dictated by preservation needs, but also by the constraints of the BnF's digital library, Gallica. Working on third-party archiving meant building bricks for pre-ingest that were as simple and as universal as possible in order to make a nonetheless acceptable SIP. The focus was thus on bit-level preservation, with an added service of metadata processing. The client can submit metadata files with its information packages, and the metadata will be mapped through an XSLT file to the descriptive metadata section of the SIP's METS file. The rules for detecting these metadata files and the content of the XSLT are entered into the service level agreement.

The idea was that after a first phase of experimenting with bit-level preservation, upgrades to the third-party archiving track would be designed in accordance with actual clients' needs and requirements.

The internal benefits of developing software for non-BnF communities were not negligible: the repository now had a redesigned pre-ingest architecture, that relied on common functions, and a model for simple, versatile pre-ingest modules which could be re-used for preserving library collections when time, resources or maturity meant that advanced pre-ingest functions could not or should not be developed. It has been used to deal with the digital versions of advertising posters, for instance.

3.2 Defining prices

There have been tensions from the start between different objectives in opening the BnF's repository to other users, and it is no surprise that they resurfaced throughout the long process of setting the prices for archiving.

[9] “Association Française de Normalisation”, that is, the French Association for Standardization.

[10] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=54911

[11] http://datasealofapproval.org.

First came the question of what the library wanted to sell: software, or a service? Around the time when SPAR was becoming operational, president Nicolas Sarkozy launch the idea of a "Great Loan", whereby the French State would borrow money to finance projects in new technologies, including the digital sector, and stave off economic crisis. As the Loan was shaping up to become the "Investissements d'avenir" (investing in the future) program in 2010, many public institutions were scrambling to set up projects that would fit what was known of the governmental action. As it was designed to boost the economy, sizeable returns on investment were expected, and the BnF sought out partners to monetize one of its big digital assets, SPAR, on a large scale. The potential partners who came forward thought of selling maintenance services around the software, to be used as a whole, or of making use of the BnF's vast digital storage facilities to bring down the costs of their services. Nothing came to fruition, although some talks are continuing along the same lines with other institutions. Yet it also meant that any ideas of making SPAR an open-source project, which would have required some initial spending and would not have yielded visible financial rewards, were put aside.

Meanwhile, the initial idea to have a track dedicated to third-party archiving within the BnF's instance of the system was not put in jeopardy by the discussions surrounding "Investissements d'avenir". It was, however, almost as difficult to price, first because the BnF, as a library, has little experience in selling goods and services, secondly, because the market for digital preservation services is still emerging, with very different offers, from cloud storage to comprehensive preservation, and not even private firms have a strong hold on their price range. The library decided to contract a consulting firm to get an idea of how much it could ask. The prices that emerged from the study then had to be validated by the ministries of Culture and Finance, who had their own priorities and policies.

Three factors were taken into account to set the prices for third-party archiving: existing pricing tiers on the market, willingness to pay, and the costs to the BnF of adding an extra terabyte of third-party data into the existing repository. Two price points were taken into account: direct costs due to the extra data (storage media, servers, manpower for ingest operations...), and global costs of maintaining the repository (software, hardware, expertise...), to be shared by the BnF and its customers.

Regarding the investment costs of preserving extra data, the consultants considered volumes, type of storage media (two tapes, two tapes and one disk, two tapes and two disk copies, and the benefits in terms of access gained with each extra disk copy), complexity of data ingest (from a self-serve dropbox to a tailored solution) and contract duration (a longer contract would level off the costs). The most recently acquired media were used as a basis to define the cost of the storage, brought back to the cost per Tb; a share of the costs of the tape libraries, tape readers and disk arrays was added. The human resources costs were assimilated to three days of an engineer’s time should the client do most of the ingestion operation, ten to twenty days for a tailored solution. The costs of developing SPAR's software were calculated for a year, then divided by the number of existing terabytes.

As for the maintenance costs, they include support for the hardware, the software, the network and the sites, as well as a proportion of the human resources costs for daily operations, and assistance to the customer when needed (one and a half days for the generic ingest process, or three days for tailored solutions). On top of that is added 17.5% of the investments costs for one customer from the fourth year on, for maintaining the material acquired specifically.

Finally, as of Spring 2013, two tiers of clients have been identified:

- clients for the archiving services only;
- cultural institutions that have a partnership with the BnF as "Pôles associés" and benefit from other services (see 3.2.2).

3.2.1 Dedicated services: BnF Archivage numérique

Regular clients for the third-party archiving services will pay according to:

- the size of their collections, per terabyte. There is a decreasing price scale for 1Tb, then 2 to 5Tb, 6 to 9Tb, 10 to 29Tb and 30 to 49Tb;
- the number of copies they want made. The standard deal is for two copies on tape, one on each of the BnF's storage sites. One or two copies on disk come at an extra charge;
- the planned duration of archiving. So far, a decreasing price scale has been set for 3, 5 and 8 years;
- the level of service. Clients using the service autonomously, more or less as a drop box, pay less than those requesting evolutions in the code to have extra preservation functions. Those developments would in theory benefit the BnF's own collections as well, and so have been moderately priced[12].

3.2.2 Integrated services: Pôles associés

The missions of the Bibliothèque nationale de France include animating a national network of libraries[13]. As such, the BnF has distributed funding, first for catalog automation and integration to the national collective catalog, then for coordinated digitization programs. It seemed natural to promote preservation of these digitized collections. Members of the partnership programs will benefit from an 80% reduction in preservation costs if they entrust the BnF with the dissemination of their digitized materials in its digital library, Gallica. (Gallica already aggregates content from several institutions, whether through OAI-PMH indexation of content, or through the BnF's digitization programs, which include some digitized books and periodicals from partners.)

In addition, the BnF is building a Cooperation Portal extranet[14], to facilitate the management of different types of collaboration by the partners. A much-needed GUI for the monitoring of information packages' ingest, storage and dissemination is in the making, and could be a model for better communication between producers, preservation experts and repository administrators within the library as well.

3.3 Defining processes

PAIMAS[15] has been around for years (since 2004 as a CCSDS standard, 2006 as ISO 20652), and is still the only official, international standard for information exchange between the producers and the Archive. Yet the BnF has had trouble matching it to its own negotiation processes, mainly because of the many departments and teams involved in making the preservation

[12] http://www.bnf.fr/documents/archivage_num_tarifs.pdf.

[13]http://www.bnf.fr/en/professionals/national_cooperation/a.creating_national_network.html.

[14] http://espacecooperation.bnf.fr.

[15] Producer-Archive Interface Methodology Abstract Standard. Cf. http://public.ccsds.org/publications/archive/651x0m1.pdf.

services work. Potential clients are either sent to the Direction of Networks and Services if they are purely archiving clients, or to the Department of Cooperation if they are partners otherwise.

Preservation experts are in different departments according to their specialties. Different teams in the IT Department are involved when there are developments to be planned, on top of production planning to be sorted out. This is why the library has had to adapt PAIMAS to an idiosyncratic version, where phases can be aggregated, or distributed across several actors.

Meanwhile, as the BnF was contemplating courting the public archives community as clients, the Central Services for French Archives (Service Inter-ministériel des Archives de France, SIAF) published a standard for the exchange of data for archiving (Standard d'échange de données pour l'archivage, SEDA[16]). It has been developed since 2006, with version 1.0 published in September 2012, and a national standard is in the works with the name MEDONA. The standard describes formally the exchanges between the different actors during the archiving and retrieval of records, and provides an XML schema to encode the transactions. It has been created to facilitate the exchange of public records, in the realm of e-administration, between the services creating the information and the services in charge of archiving public data. Therefore its use is highly recommended to candidates seeking to sell short-term and mid-term preservation services of public archives. But the recent and rapid evolutions of the standard have led the BnF to put its implementation within the repository on hold, at least until the second semester of 2013.

4. CONCLUSION: WHAT'S NEXT?

Offers and prices for third-party archiving at the BnF are finally stabilized, with two tiers of clients, and this seems well positioned to benefit the library, through the incentive to develop new functionalities, and through some return on investment.

A first client, the Virtual Center of the National Museum of Modern Art, has led the way in taking up the offer, and this experience has helped streamline pricing, exchange processes, and workload management.

But how stable is the offer, really? The volatility of policies at the library and the state level carries an important risk at the management level. The existing clients' and partners' collections will be looked after according to contract, but what about the day-to-day operations' burden on the library's staff and resources? It is yet unclear whether the profits generated by the services will be enough to absorb the extra work, whether in setting up the administrative details of the contracts, dealing with the ingest and dissemination flows or adding new features to the repository, while maintaining an appropriate level of service for the BnF's own digital collections.

Moreover, the trend towards collaboration and sharing of resources is still being felt, as new projects emerge while budgets shrink. It is not clear at this stage which will prevail: the creation of multiple small repositories arising from the differences in size and constraints of the communities, even within the public sector, or the wish to regroup and save, and to share technology that is championed by its designers. Will the cost models be sustainable and guarantee the preservation of the partners' as well as the library's own collections?

Feedback from similar projects would help the management, as well as the team designing the software and the storage, assess its third-party archiving policy. Additional international benchmarking initiatives would benefit communities in a similar situation.

[16] http://www.archivesdefrance.culture.gouv.fr/seda/ (in French). The English presentation dates from the 2006, 0.1 version of the standard: http://www.archivesdefrance.culture.gouv.fr/seda/documentation/archives_echanges_v0-1_description_standard_v1-0-english.pdf.