Date : 01/07/2008

Digital preservation at the National Library of France: a technical and organizational overview

Emmanuelle Bermes
Head of Prospective and documentary services, Bibliographic and Digital Information Department

Isabelle Dussert Carbone
Head of the Preservation and Conservation Department

Thomas Ledoux
Project manager, IT Department

Christian Lupovici
Head of the Bibliographic and Digital Information Department

Meeting: **84. Preservation and Conservation, (PAC), Information Technology, IFLA-CDNL Alliance for Bibliographic Standards (ICABS) and Law Libraries**
Simultaneous Interpretation: Not available

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

National libraries are going digital: acceleration of the growth of digitized collections, change in the form and scope of legal deposit, development of Web archiving and records management, are placing the library in front of new challenges, including the issue of long-term preservation of this new material. This change of landscape is a technical and organizational challenge; it is a need for rethinking the library’s mission, in terms of infrastructure, project management, human resources, day-to-day activities, and so on.
The National Library of France (BnF) is in the process of creating its own digital preservation repository: SPAR (Système de Préservation et d’Archive Réparti / Distributed Archiving and Preservation System), a trusted infrastructure that will be able to collect, store, preserve and disseminate a large amount of diverse digital material, either digitized or born digital.

Technical implementation

SPAR is a trusted repository, compliant with the OAIS standard (Open Archival Information System – ISO 14721). Its design is modular and scalable. It provides a full mirroring system with monitoring and alerting capabilities along with a disaster recovery plan.

The SPAR system interacts with various production applications corresponding to the production of digital objects in the library, including digitization processes, born-digital

objects creation (like the records management), or born-digital objects harvesting in the context of Web archiving.
On the access side, these digital objects are provided by SPAR to dissemination applications which take in charge the presentation of the digital objects to end-users. These dissemination applications include the digital library of the BnF, Gallica[1], and others.

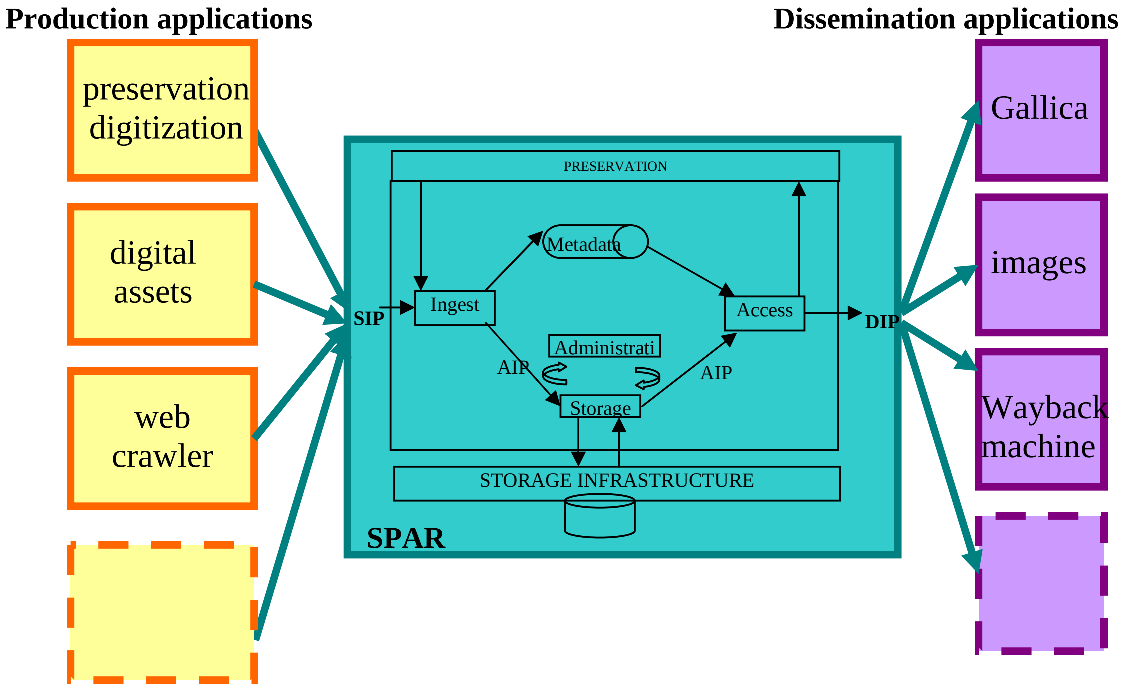

Fig. 1 - Overview of the SPAR system – functional view

The implementation of the SPAR system is decomposed in two sub-projects:

- The “SPAR Infrastructure” sub-project aims to acquire and implement the technical infrastructure needed by the system.
- The “SPAR Realization” sub-project aims to build the system in order to store, to maintain in the long-term and to communicate digital documents of the library using the available infrastructure.

Software architecture

Hence, SPAR is a long-term preservation system for digital material. Its conception is **modular** and extensible, articulated through the following generic modules:

- Ingest module
- Storage module
- Data management module
- Rights management module
- Access module
- Administration module
- Preservation module
- Technical module (« Storage Abstraction Service »)

with the addition of more specific ones:

[1] http://gallica.bnf.fr

- Pre-ingest (building of ingest packages)
- Delivery of dissemination packages.

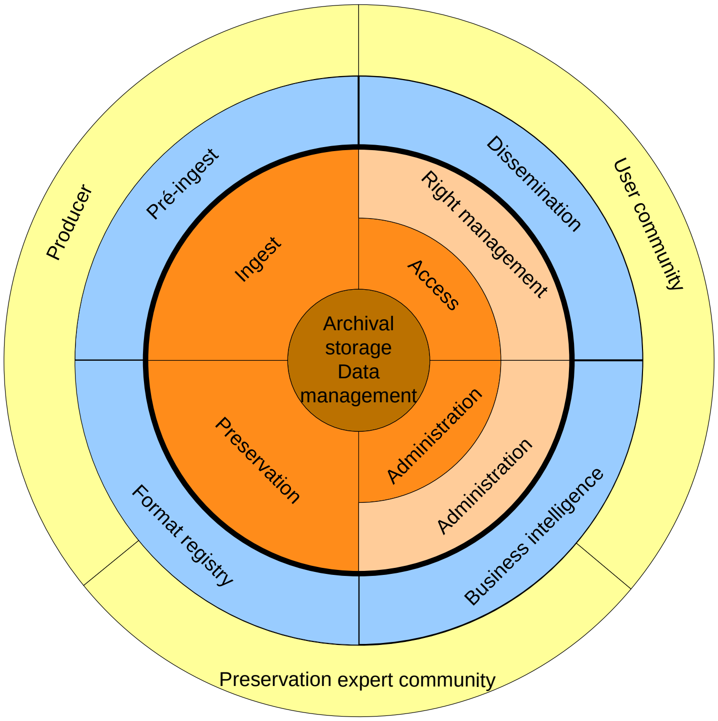

Fig. 2 - Overview of the SPAR system – conceptual view

The specificity of these two last modules comes directly from the notion of **channel**. A channel is characterized by the relations between the production or the use of the digital objects (needs and requirements) and the archival system (service agreement).
Material expected to be ingested in the future system is classified among 7 channels:

- preservation digitization: material digitized by the library from analog material for preservation purposes
- reproduction digitization: material digitized by the library from analog material on an on-demand basis (patrons ordering)
- automated legal deposit: Web archiving collected through automated processes (crawlers)
- negotiated legal deposit: material ingested for legal deposit purposes, collected through particular processes to be negotiated with producers
- records management: material created in a digital form by the library during its own activities
- third party archiving: material archived for the account of a third party
- acquisitions: material acquired on a fee-based basis or collected through gifts, legacy etc.

Infrastructure

The technical module “Storage Abstraction Service” guarantees the independence from the physical infrastructure by ways of **storage units**. Those units are defined by a set of characteristics and services which allows the match with a particular storage component or more usually a set of those elements (for example, the association of a pool of tapes with a partition in a disk array).These different units are defined by the administrators using the storage components they have to match at best the requirements defined by the policies.

The actual infrastructure is made of the following components:

- the infrastructure is distributed along 2 geographically distinct sites,
- each site owns is own tape library used for hosting the AIP,
- the exchanges (deposit or delivery) are made through disk arrays called SSS (for the entrance) and SSC (for the dissemination);
- management (SUG) and treatment (SUT) servers host the software,
- media supervision components provide monitoring and alerting capabilities.

Realization

The development of this system is planned through a 4 year contract. The first year aims to realize the common core of the system as well as the channel for the preservation digitization. The other channels will be developed in an iterative mode.

The overall system is built using reusable components assembled with an Open Source framework: FedoraCommons. This framework brings the benefit of a mature development, guarantee of reliability, and sustainability as it is associated with a wide community.

Information model

However, the technical infrastructure is not sufficient to achieve the goal of long term preservation: a strong commitment is made within the SPAR project for persistent data management and organizational viability in order to ensure that the system will not become obsolete or suffer a lack of consistent day-to-day management in the future.
On the data management point of view, all the conception is based on authoritative standards such as METS (Metadata Encoding & Transmission Standard) and PREMIS (Preservation Metadata Maintenance Activity).

Metadata standards

The METS standard has been chosen as a packaging format. It binds together the data objects (digital files or bit streams as managed by the Fedora system) and the metadata. The latter includes a minimal subset of descriptive metadata in Dublin Core format, imported from the bibliographic catalog through the OAI-PMH repository; but the main part of the metadata is administrative and technical metadata generated by the system itself. The provenance metadata, ensuring the audit trail of changes that occur within the system, are recorded as a series of events encoded according to the PREMIS standard. The technical metadata is extracted from the digital files during the ingest process, and encoded in the METS technical metadata section using appropriate extension schemas such as MIX for still images and TextMD for textual documents. The METS manifest thus obtained provides a global view of the digital object to be preserved, and is stored along with the digital files in the preservation system. However, to ensure the accessibility of this metadata and the ability to query it in a flexible way, an innovative data management system will be developed. The data management will be based on a mapping from METS to RDF and the resulting RDF data will be stored in a RDF triple store. This technology will provide the most accurate access to all information stored in the system and will allow a strong monitoring of the digital objects in the perspective of preservation strategies such as migration and emulation.

Managing granularity, versioning and preservation strategies

Along with this metadata model, a generic information model for granularity and versioning management has been designed. This model includes four levels of granularity (set – group – object – file) which cover all the different possible configurations of complex objects, the “set” level being recursive. In the METS manifest, this structure is defined in the structural map section:

- **set**: this level is used for a collection of digital objects, the title of a serial or a multimedia document
- **group**: this level corresponds to the digital object : for instance, a monograph, an issue of a serial, a film
- **object**: a consistent part of the digital object, such as an image, a page, a track for an audio CD
- **file**: the data object (digital file or bit stream).

For instance, an issue of a digitized newspaper is considered as a “group”. The “set” corresponds to the title of the newspaper. Each page of the issue constitutes an “object”, itself composed of two “files”: one for the image of the page and one for the OCR transcription. The “set” and “group” levels are both considered as independent, loosely coupled packages with their own METS manifest. The different types of metadata are applied to relevant levels.

One challenge was to find a balance between the necessity of updating digital objects, and the security and reliability of the archive. Thus, it was necessary to allow regular updates of metadata (including descriptive ones) and progressive amelioration of the data objects (for instance addition of a new version of the OCR file when the OCR accuracy is improved due to state of the art evolution). The versioning system takes this into account and provides a flexible lifecycle management based on versions and editions. When the data objects are affected by the update (ex. replacing or deleting a digital file) a new version is created and the older version is preserved. On the contrary, when only the metadata is changed, or data objects added but neither modified nor deleted, a new edition is created and replaces the previous state of the digital object.

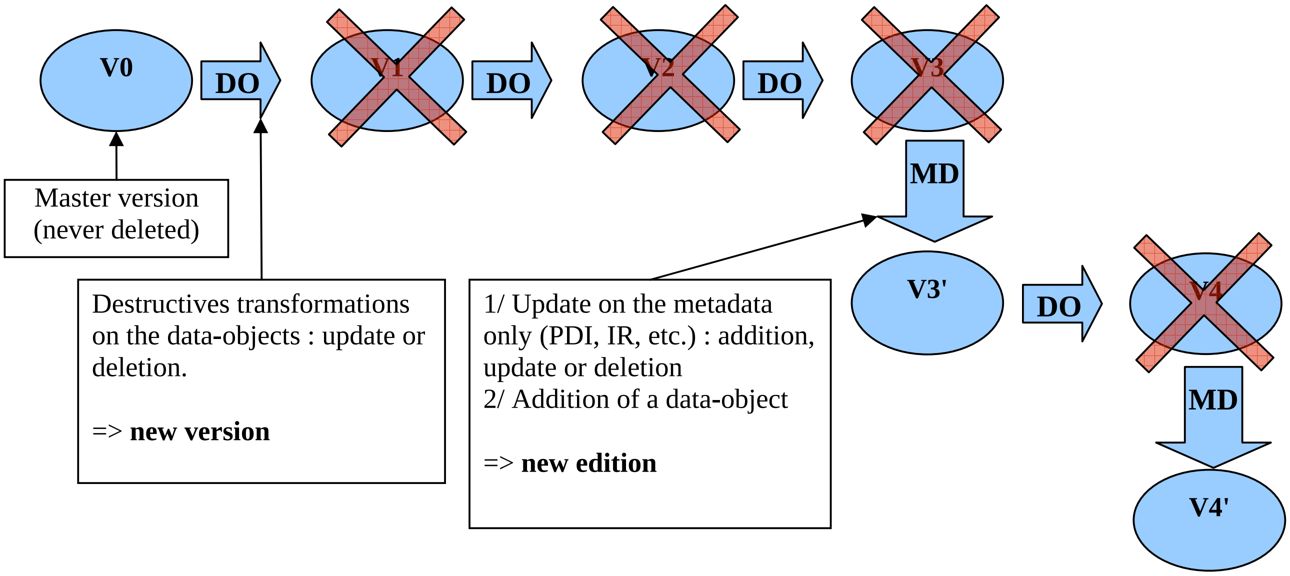
V0
DO
V1
DO
V2
DO
V3
MD
V3'
DO
V4
MD
V4'
Master version (never deleted)
Destructives transformations on the data-objects : update or deletion.
=> **new version**
1/ Update on the metadata only (PDI, IR, etc.) : addition, update or deletion
2/ Addition of a data-object
=> **new edition**

Fig. 3 – Lifecycle management for versions and editions of packages

Thanks to this model, we expect that future preservation strategies will be managed in a flexible way. Operations that present a risk of data loss will be secured by the versioning management; however, the system will not be paralysed by this security commitment.

Digital collection management

Since the SPAR system seeks to reach a high level of modularity and flexibility, a strong commitment for good digital collection management is necessary. This will be achieved through a set of formalized procedures that govern the relationship between producer and archive. This negotiation is based on the PAIMAS standard and on service level agreements. For each channel, the producer and the archive will negotiate three types of policies: one for ingest, one for preservation and one for dissemination. These policies, or service level agreements, are formalised procedures which will force the producer to express his needs in quantifiable terms that are then to be converted to formal rules to be used by the system. The policies are to be preserved within the system as well as the objects themselves, so that the system is completely self-described.

The service level agreement is a formal document that describes in an extensive way the processes, actors, content and strategies associated with a channel. The detailed technical notebook describes precisely the origin of the metadata and the structure of the granularity of packages. The service level agreement is composed of three policies for each channel:

- the ingest policy (acceptable formats, volume, security levels, …),
- the preservation policy (retention time, assurance levels, …),
- the access policy (dissemination formats, time, availability, …).

The ingest policy allows the validation of the producer's ingests and the taking on the responsibilities according to the category of format:

Code	Format category	Description
00	Stored	Format which technical characteristics are unknown (not identified) and for which only bit-stream preservation is ensured.
01	Identified	Format which technical characteristics are known (identified through a format registry) but for which no migration or emulation path is envisioned. An identified format becomes managed or known if such path is implemented.
10	Known	Identified format for which the BnF owns at least one reference tool, knows its uses, keeps track of its evolution and for which the BnF has defined a path either to transform it in a managed format or to emulate it.
11	Managed	Known format for which BnF owns the published documentation and at least one reference tool, keeps track of its evolution and for which the BnF has defined specific constraints with the producers.

The preservation policy defines where the archival information packages (AIP) are stored and how their lifecycle is managed. A managed format is required for a migration strategy, while a known format is required for emulation.

The access policy defines in particular the dissemination formats as well as the access constraints which determine if the disseminated objects must be pre-calculated or generated on the fly.

The service level agreement also takes into account a risk management strategy in order to ensure the transfer of responsibility from the producer to the archive. The overall process is a guarantee of good knowledge of the risks that lay upon the digital objects, and appropriate commitments of the producer to ingest appropriate material and of the archive to take all the necessary actions in order to provide the preservation service.

Organizational issues

On the organizational point of view, specific library activities are being developed: metadata management, document analysis, collection management, policies definition, rights management, etc.

Mass digitization and preservation purposes

These new specific activities implied by the development of SPAR are not only IT activities but also collection management. Access facilities given by digitization and OCR conversion significantly improve the comfort and user experience compared to the use of paper or microfilm material. For this reason, the demand for digitization programs is increasing, coming from all the departments in charge of collection management. Digitization is now also envisioned for preservation matters, the digitized object being considered as a new type of surrogate, able to prevent the original from suffering degradations due to public use. Last but not least, digitization is undertaken for intellectual purposes, in the case of web thematic portals and virtual collections. These different motivations for digitization sometimes lead to confusion when preparing mass digitization projects: the scientific and intellectual policy, including specific thematic presentations of collections, is sometimes opposed to the preservation digitization policy.

The work on SPAR organization has developed confidence in digital preservation, so that the library is in the process of giving up microforms in favor of digital surrogates. This orientation has to be financially planned. In the analog world, the cost of stacks and traditional conservation operations on collections are planned and calculated; in the digital world, the library also has to forecast and project the costs, including:

- storage and preservation for a huge amount of digital material,
- software development and maintenance,
- hardware upgrades and maintenance.

These projections are similar to the building of a material library, and they present very similar needs:

Analog world	Digital world
building	computer room
stacks	storage
OPAC	search engine
management system	administration
restoration	preservation system
call number	identifier
reservation and loan	dissemination systems

It remains unsure today that the library actually evaluated the real costs of such an ambitious program. In addition, a set of specific preservation activities is required, as strongly as the technical aspects, as a factor of success in a long term digital preservation program.

Preservation activities

Digital preservation is a true preservation matter; therefore well known methodologies such as risk assessment strategies, can be applied to digital material. The preservation strategies are defined in programs including preventive and curative measures, in order to remain at a good technological level, and ensure refreshing, migration, replication, emulation. Regular review of the digital collection is the only way to check that the content ingested in the SPAR system remains unaltered, and this monitoring activity is also well known of preservation experts. The preservation metadata represent the equivalent of restoration information that is collected in preservation actions for analog material, and they guarantee the possibility to follow the digital material history. Global disaster-recovery plans for the library have to include digital preservation systems. The specific activities of negotiation between the producer and archive are well known of restoration services, and they have to be extended to the discussion on digital material.

For born-digital material such as Web archives and other kinds of digital medias (CDs, DVDs, video games…) the challenge for preservation is even greater. On the technical point of view, this type of material can raise specific issues such as specific hardware and software environment, DRM (digital rights management) protections, rare or unknown file formats… which tend to make it more difficult to create the conditions for long term preservation. Moreover, this material, in particular Web archives, is subject to a change of scale: the masses of collected items require high storage capacity and involve a selection and sampling approach which is not familiar to librarians in the context of legal deposit and preservation. This approach might seem rather similar to archivists' practices, and suppose an effort to

inform the future user communities regarding the state of collections and technical limits that apply to the harvesting process.
All these specific preservation activities (preservation planning, collections monitoring, metadata creation and querying, recovery plans, producer-archive negotiation) emphasize the need to identify the responsibilities on the evolution and day-to-day management of SPAR. The expertise needed to assess particular collections in order to make choices, define priorities and plan preservation actions, are traditional librarian skills which need to be adapted to the digital environment. This requires specific training for the staff in order to develop skills in digital preservation.

Collaboration in the realm of digital preservation

SPAR is an ambitious project which the BnF doesn't intend to achieve alone. Worldwide collaboration is considered as an essential part of the work, and it includes an active participation within open source communities (Fedora and others), or setting up interests groups like PASIG (Preservation and Archiving Special Interest Group, in collaboration with Sun Microsystems).
Collaboration through specific groups like PIN (Pérennisation de l'Information numérique, expertise and dissemination interest group in France) or IIPC (International Internet Preservation Consortium) is essential in the case of digital objects preservation, to raise awareness or to define best practices and to bring technical and organizational solutions we all need while facing the digital preservation challenge.

References

Bermès Emmanuelle, "Risk Management: methodological principles" in *International preservation news*, n°41, july 2007.

Calderan, Lisette (dir) et al. *Pérenniser le document numérique. Séminaire INRIA, 2-6 octobre 2006, Amboise.* ADBS, 2006.

Fedora commons. http://www.fedora-commons.org/

Kaczmarek Joanne et al. "Using the Audit Checklist for the Certification of a Trusted Digital Repository as a Framework for Evaluating Repository Software Applications. A Progress Report." *D-Lib Magazine*, Volume 12 Number 12, December 2006.

Kolding Nielsen, Erland. "Digitisation of library material in Europe: problems, obstacles and perspectives anno 2007." *Library quaterly*, Vol 18, 2008, n°1

Lavoie Brian, Gartner, Richard. *Preservation metadata. Technology watch report*. OCLC/DPC, september 2005.

METS. *Metadata encoding and transmission standard*. http://www.loc.gov/standards/mets/

MIX. *NISO metadata for images in XML format*. http://www.loc.gov/standards/mix/

PREMIS. *Preservation metadata*. http://www.loc.gov/standards/premis/

Producer-Archive methodology abstract standard. CCSDS, Blue book, may 2004. http://public.ccsds.org/publications/archive/651x0b1.pdf

Reference model for an open archival information system (OAIS). CCSDS, Blue book, January 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf

Rieger Oya Y., *Preservation in the Age of Large-Scale Digitization. A White Paper*. CLIR, February, 2008.

TextMD. Technical metadata for text. http://www.loc.gov/standards/textMD/

Verheul Ingebord, *Networking for digital preservation Current practice in 15 national libraries*. KB/IFLA/Saur, 2006

Verheusen, Astrid. “Mass digitisation by libraries : issues concerning organisation, quality and efficiency.” *Library quaterly*, Vol 18, 2008, n°1