Experiment, Document & Decide: a Collaborative Approach to Preservation Planning at the BnF

Bertrand Caron
Department of Bibliographic and Digital Information
Bibliothèque nationale de France
bertrand.caron@bnf.fr

Thomas Ledoux
Department of Information Technology
Bibliothèque nationale de France
thomas.ledoux@bnf.fr

Stéphane Reecht
Department of Conservation and Preservation
Bibliothèque nationale de France
stephane.reecht@bnf.fr

Jean-Philippe Tramoni
Department of Information Technology
Bibliothèque nationale de France
jean-philippe.tramoni@bnf.fr

ABSTRACT

The National Library of France (BnF) has recently implemented a new module for its Scalable Preservation and Archiving Repository (SPAR) to set up preservation strategies based on formats, agents, workflows, tools and tests, and managed as reference packages in the Archive.

This module aims to fulfill an objective: for SPAR to be fully self-documented. Formats, agents and workflows are formally described and preserved along with the Information packages in which such elements are involved. Although this was a feature that was included from the beginnings of SPAR, the new Preservation Planning module aims to provide a tool that can more easily build these reference packages and that will more closely involve domain experts and the IT department in the processes of preservation planning. But the main innovation lies in the documentation of decisions that directed their selection as standards in SPAR: test data are now preserved as a new kind of reference package.

General Terms

Preservation strategies and workflows; innovative practice.

Keywords

Preservation planning, decision documentation, community involvement.

1. INTRODUCTION

Since the operational launch of SPAR in 2010, the BnF has had to face a growing diversity of digital documents (heritage digitization in 2010, third-party archiving in 2011, web archiving in 2013, legal deposit of ebooks planned for 2015). Ingest and preservation of these specific materials led the BnF to implement many different workflows involving characterization, processing and transformation tools.

The BnF felt the urge to record the choices made about these operations not only within the system logs but also within the repository itself. From data objects on which tests were performed to results of said tests using a software tool, every step explaining the decisions that led experts to carry out a specific preservation plan has to be preserved.

Following the path initiated by the experimental tool Plato [1] and based on discussions with various communities, the "Preservation planning" module was developed to address this specific need. Although all activities of the Preservation planning OAIS entity were not taken over, its first version allows experts to develop preservation strategies and standards and keep track of their elaboration.

The module is provided with a user-friendly interface and several levels of authorization; its objective is to foster collaborative work between experts from different departments of the library.

2. WHY A PRESERVATION PLANNING MODULE?

2.1 What Does OAIS Say?

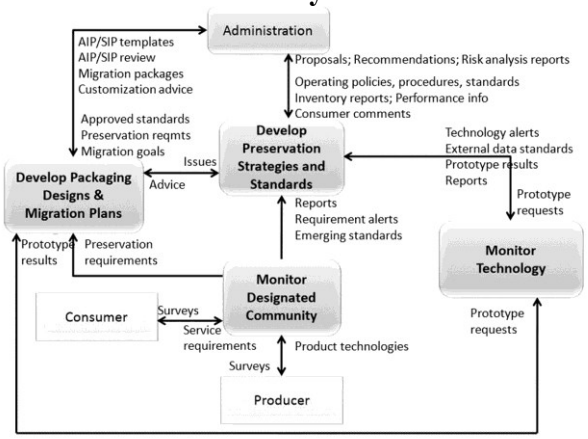

Figure 1. Functions of the 'Preservation planning' entity (source: Reference model for an Open Archival Information System (OAIS) [2])

As defined by the OAIS standard, the Preservation Planning Functional Entity "provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the Long

Term, even if the original computing environment becomes obsolete."

Monitoring the environment of the archive is achieved through two functions: Monitor Designated Community and Monitor Technology. The former calls for interactions with members of the community in order to track changes in their service requirements or product technologies. The latter requires performing surveillance on emerging standards or technologies. Any changes are reported to the two other functions of this entity which are responsible for defining, developing and validating preservation plans and appropriate tools (see [Figure 1]).

The Develop Preservation Strategies and Standards function "is responsible for developing and recommending strategies and standards, and for assessing risks, to enable the Archive to make informed tradeoffs as it establishes standards, sets policies, and manages its system infrastructure." In response to the reports about identified changes in the environment of the archive, this function will have to estimate possible updates in archive operations, including policies, procedures, standards and tools. This evaluation may require prototyping and testing of these updates such as: SIP/AIP templates, submission requirements, new or modified file formats and tools for identifying and characterizing these formats. This process enables the Develop Preservation Strategies and Standards function to issue recommendations and advice to deal with the incoming changes.

Carrying out these recommendations is the responsibility of the fourth function. The Develop Packaging Designs and Migration Plans function "develops new Information Package designs and detailed migration plans and prototypes, to implement Administration policies and directives." This task will include development of new AIP designs, prototype software, test plans, community review plans and implementation plans for phasing in the new AIPs, and may call on expertise or resources from other functions. After proper testing and validation, the developed elements – plans, AIP designs and templates, software – will be sent as a package to be put into production.

2.2 Context of SPAR

SPAR (Scalable Preservation and Archiving Repository) is the BnF preservation system, compliant with the OAIS model. Its scope is to manage all entities that can be automated through modules corresponding to the OAIS entities.

2.2.1 Tracks and Channels

In SPAR, sets of documents to be ingested are processed by tracks and channels (sub-tracks), according to their nature (e.g., digitized books, audiovisual files, web archives, administrative records), their legal framework, and the way the BnF plans to manage their life cycle and apply preservation strategies. At the present time, SPAR ingests objects through six tracks: digitized documents and associated files, audiovisual objects, web legal deposit (ARC or WARC files), negotiated legal deposit (ebooks), administrative records, and third-party archiving (various kinds of files, from partners outside the institution); several others are in progress.

Every channel is managed by Service Level Agreements (SLAs), negotiated between the Producer and the Archive. They define the terms of ingest, preservation and dissemination (e.g., formats accepted, maximum size of packages, availability of service). Each SLA is transcribed in XML files that configure the system.

2.2.2 Reference Packages

SPAR is a self-documented system. It holds and preserves its own reference packages, ingested in a Reference track. These packages describe and identify every component of the preservation policy: formats, agents (software products, modules, processes, and humans), ontologies, classification systems, tracks and channels. SPAR uses them to document every process and, because most of them are machine-actionable, to perform automatic operations (e.g., checks, extractions, transformations). Thus, part of Representation Information and Preservation Description Information is preserved as well as the Data Objects, allowing reference to common information in every package through links and unique identifiers. This way, the verbosity of the manifests is reduced.

2.2.3 Actors

Many OAIS activities and functions cannot be fully automated and need human actors: administrators, preservation experts, developers, risk managers, collection managers and track managers. The Library has taken the measure of the challenge and is working on its organization (see [3], [4] and [5]).

Administrators are members of the IT department. They are responsible for deployment of new software versions, channels, requirements, etc. and, above all, for ensuring the system meets the SLAs in the daily production.

Preservation experts are members of many departments (e.g., from bibliographic information, IT, digitization departments, etc.); their expertise is functional or technical (on formats, storage, technical or bibliographic metadata, etc.). They are involved in standards elaboration. They are the core team that monitors the evolution of technology and the needs.

Track managers are members of departments who receive digital material to be preserved in SPAR: legal deposit, preservation, archiving mission, etc. They monitor a specific producer community and are responsible for ingest and preservation of documents belonging to their track.

2.3 The Preservation Planning Module: a Bottom-up Strategy

Up to now, development of SPAR was mainly concentrated on Ingest, Storage, Data management and Administration modules.

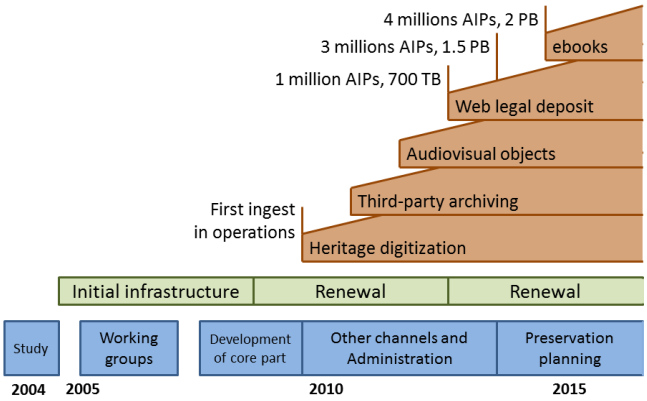

Figure 2. SPAR milestones

In 2014, the BnF decided to develop a module intended to fulfill the functions and activities of another OAIS entity: Preservation Planning.

2.3.1 Building Reference Packages

Formerly, the reference packages were discussed between project stakeholders then coded by developers in XML. Now they can be produced by the interface and modified at any time by a larger

community of allowed users. Indeed, the new module is designed to foster collaborative work between administrators, track managers and preservation experts on the reference package elaboration.

In general, some of the benefits expected are:

- Greater speed and reactivity, involving common expertise throughout the library;
- Increased trustworthiness, thanks to a validation workflow involving more people; and
- Increased visibility of preservation activities.

2.3.2 Documenting Decisions about Standards

Whereas formats, channels and agents had always been preserved in SPAR, the Preservation planning module brings a whole new functionality. The entire decision process, from basic migration or characterization tests to preservation plans on a large scale, is now documented in reference packages ingested in SPAR.

Two major cases are foreseen:

1) Characterization. An upcoming file format has to be preserved in SPAR. What characterization tool will be used? What technical metadata is needed? See use case below, 4.1.
2) Migration. The new file has to be transformed when ingested into another format, preferably an open one. Which final format will be chosen? Which transformation tool will be used? See use case below, 4.2.

In the end, such tests will result in a new SLA with a new definition of ingest settings.

Four package types were defined to document this decision process. First, the Data Objects on which tests are carried out (initial "test data") and, in the case of a migration, the result of said tests (transformed "test data") are preserved. Secondly, characterization of initial and transformed data is preserved in "test metadata" packages referring to the Data Objects processed and to the used tool, described and ingested as an agent in the reference channel. One or more "test campaigns" are performed out of a significant number of tests leading to a decision that is implemented in a "preservation plan".

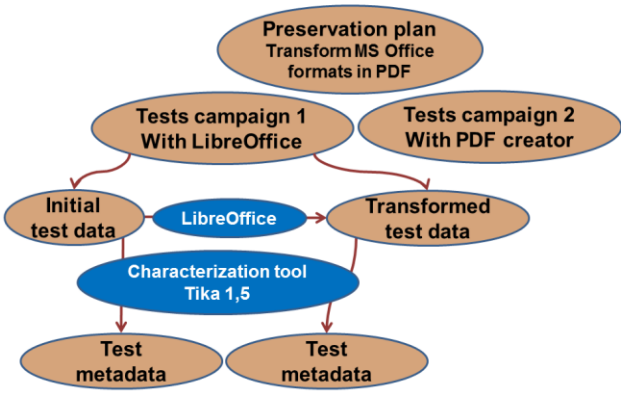

Figure 3. General organization of test packages

All this information is meant to be preserved in order to document the performed experiments, to give the material to allow the reproduction of these experiments, and to have a stable decision base in order to come back later and be able to reconsider such decisions.

3. IMPLEMENTATION CHOICES

3.1 A New Approval Workflow

In order to formalize the decision process, a validation circuit was organized, and several levels of authorization were defined according to it.

The following steps of the process were determined:

1) A specific need is submitted by track managers (e.g. a new file format to be preserved) or by preservation experts (e.g. a risk of obsolescence is identified for a specific format).
2) Tests are carried out locally by preservation experts to solve the issue (e.g. characterization and migration tools are run on sample documents).
3) If there has been a transformation, the transformed files are characterized by suitable software.
4) Sample documents and transformed documents are ingested in the Archive. Results of characterization are ingested as well.
5) A decision is taken about how to address the issue, given the tests results. Note that currently no method is defined to come up with the decision: anything such as mind maps, SWOT analysis or decision matrices can be used.
6) Preservation experts create new SLAs that take into account the new preservation strategy (the file format will be characterized or transformed into another format by a specific tool when ingested).
7) Programmers develop a technical solution (e.g. implementation of new characterization tools) and test it.
8) When ready, the technical solution is activated by the Administration.

3.2 General Architecture

The module architecture had to reflect the current organization of SPAR. At the same time the module was developed, a working group raised some important organizational issues about the role and attributions of every human agent related to SPAR. The Preservation planning module reflected these changes.

Managing several levels of authorization was a particularly important point in the module, as it gave direct capacity to SPAR's settings to agents out of the Administration module. Track managers have rights limited to definition of channels, whereas administrators can read and modify every type of package. Format experts have no rights on the channel packages but have a key role in the elaboration of formats and agent packages.

As the needs of the actors in SPAR are different and might be conflicting, multiple instances of the system have been installed.

- A validation platform is used by developers and the Quality Assurance (QA) team to build and test new versions of the software and to validate it.
- A sandbox platform is used by preservation experts and track managers to elaborate reference packages. This instance is also used for training and is regularly cleaned up.

- The production platform holds SPAR's current deployed version.

These instances are used in the approval workflow described above as follows:

1) The reference packages are elaborated collaboratively on the sandbox platform.
2) When ready, they are transferred to the QA team who tests them on the validation platform.
3) In the end, they are activated by the Administration and ingested in the production platform.

Going back to the Preservation Planning OAIS Entity, the SPAR implementation can be summarized as below:

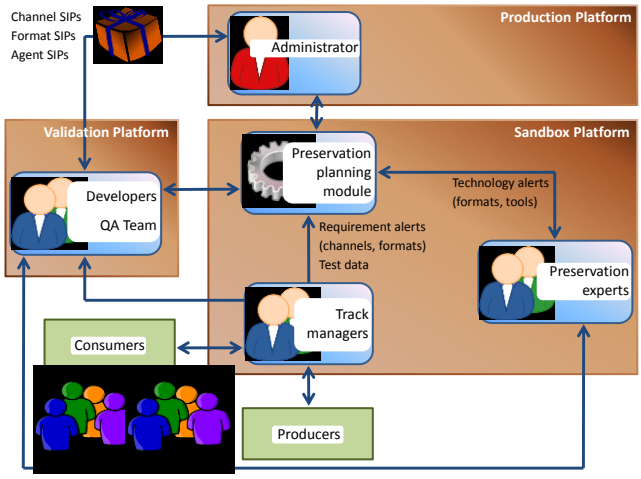

Figure 4. SPAR implementation of the Preservation Planning OAIS Entity

3.3 Ergonomics and Functionalities

Technically, the module aims to help building step by step a complete reference SIP and to transfer it into the repository. A previously ingested reference package can be updated with new requirements, thus creating a new version of this package.

Updating or creating a new package from an existing one is now easier: authorized users can ask for retrieval of an entire reference package, copy it, modify only the relevant information and ingest it again. The system delivers into a user-specific folder the manifest and the Data Objects contained in the requested package.

The interface usability was a challenge, as it was meant to bridge a gap between domain experts and IT staff. Vocabulary used in the interfaces had to be clear, consistent, precise and, preferably, agnostic about specific metadata terms.

In order to produce a complete machine-actionable reference package, different interfaces are displayed sequentially and cannot be accessed if the required information is not provided at each step. Compliance checks are carried out when moving from one interface to the next.

Common templates to several types of reference packages are defined to associate files, define events occurred before the package ingestion or enter its descriptive metadata.

The information provided by the user is recorded within the SIP manifest or within data files in which channel, software or format significant properties are stated. The preservation policy of the Reference track specifies that every version of the packages ingested by the track must be indefinitely preserved. In this way, one can always refer to requirements in effect at the time any event of package ingestion, preservation or dissemination occurred.

3.4 An Example: How to Create a Channel Reference Package?

Every channel in the SPAR repository has its own ingest requirements, preservation strategies and dissemination conditions.

The module allows track managers to create new channels and modify existing ones by updating the Service Level Agreements. This has immediate effects on packages ingest, preservation and dissemination.

The interfaces provide a set of information like patterns for descriptive metadata detection, possible transformations of input files, different files groups and formats allowed for each one, files minimum and maximum size, frequency of fixity checks, storage location, and documentation about the channel.

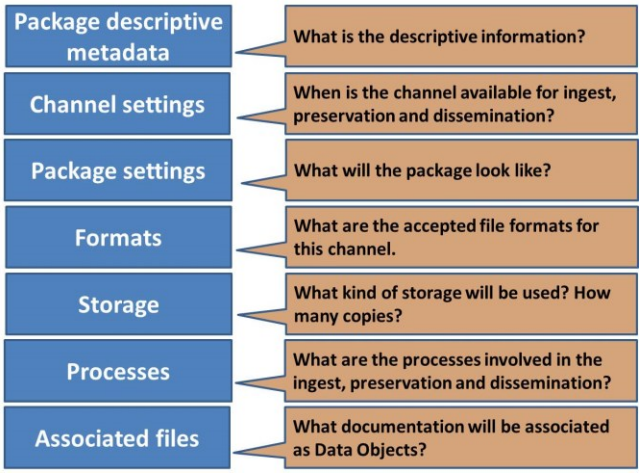

Figure 5. Channel reference package elaboration sequence

At the end of the process, the reference package contains a METS manifest, a complete description of the channel, and the three SLAs (concerning packages ingest, preservation and dissemination) expressed in XML. Three associated Schematron files are used to check the manifests of every package submitted to the channel during its lifecycle.

4. SOME REAL USE CASES

4.1 A New Format for Heritage Digitization: JPEG 2000

The BnF's digitization program was primarily focused on producing images in uncompressed TIFF v6 format which is the preferential preservation format in this case. Due to the increasing volume of data (more than 1 PB), the switch to a compressed format was required by the track manager.

Thanks to the collaboration with other heritage institutions, the choice of the JPEG2000 format was appealing [6]. In order to determine which exact settings the Library should require for such a format, a set of sample TIFF images digitized from a vast diversity of material was assembled in a reference tests package. This package was then transformed using the kakadu tool with various settings and the results were compared in order to define the acceptable compression ratio in a similar fashion as described in [7]. In parallel, we use the jpylyzer tool [8] and an XSLT

transformation to generate the corresponding test metadata package in the MIX format. We then had a way to ensure that the new images kept the significant properties of the images while taking less space.

Once the different settings were selected, the Digitized Program track manager was able to modify the reference package of its channel and the preservation system was able to ingest JPEG2000 files, characterized with the adequate tool.

4.2 Transforming Office Documents to PDF

In the course of elaborating the 'Administrative Records' track, a need for an additional preserved copy in PDF format of all the office documents became apparent.

Once again, a set of files were sampled from our production databases trying to target a large time period as well as to vary the versions of production software. As shown in [Figure 3], various tools to make the transformation were tried.

The question of finding a common format to represent the technical metadata led BnF to XMP [9], as the only one applicable for such a diversity of formats. The use of the Tika tool [10] to generate the test metadata packages provides a way to evaluate the well-formedness of the output as well as to compare the different outputs.

Currently, we have discovered that no tool is efficient enough to ensure a perfect transformation to PDF; such a conclusion reinforces the strategy of keeping both representations of the files (the original one and the transformed one) in the Archive.

5. CONCLUSION

As the module has been implemented recently, the BnF has little feedback from its potential users. Appropriation and community involvement will raise new issues and should be addressed in another paper in the years ahead.

However, using this module to elaborate in common SPAR standards has already shown good results, improving interaction between domain experts and IT members, and quality. But the module is one among other results of the BnF's will to involve more closely librarians in their digital collections preservation.

Finally, the module is likely to undergo evolution, as preservation planning encompasses many more aspects than only creating reference packages. Among them, being able to perform tests or migrations with tools stored in SPAR directly from the planning preservation module is foreseen.

6. ACKNOWLEDGMENTS

The authors would like to thank all the members of the SPAR team for their commitment and their invaluable contributions. The Open Preservation Foundation and all its members were paramount in sharing practices and maturation of ideas.

7. REFERENCES

[1] Becker, C., Kulovits, H., Rauber, A., and Hofman, H. 2008. Plato: a service-oriented decision support system for preservation planning. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (JCDL '08) (Pittsburgh, Pennsylvania, June 16-20, 2008). DOI= http://doi.acm.org/10.1145/1378889.1378954

[2] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System*. 2012. CCSDS 650.0-M-2. http://public.ccsds.org/publications/archive/650x0m2.pdf

[3] Bermès, E., and Fauduet, L. 2011. The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France. In *International Journal of Digital Curation*, vol. 6, no. 1, 226-237. http://www.ijdc.net/index.php/ijdc/article/view/175/244

[4] Derrot S., Fauduet L., Oury C., and Peyrard S. 2012. Preservation is Knowledge: A community-driven preservation approach. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012). https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf

[5] Clatin, M., Fauduet, L., Oury, C., and Tramoni, J. P. 2014. Digital curators at work: analyzing emerging professional identities at the Bibliothèque nationale de France (BnF). In *IFLA World Library and Information Congress* (Lyon, France, August 2014). https://hal-bnf.archives-ouvertes.fr/hal-01098526

[6] Buckley, R. 2013. Using lossy jpeg 2000 compression for archival master files. http://www.digitizationguidelines.gov/still-image/documents/JP2LossyCompression.pdf

[7] Martin, S., and Macleod, M. 2013. Analysis of the variability in digitised images compared to the distortion introduced by compression. In *Proceedings of the 10th International Conference on Preservation of Digital Objects* (Lisbon, Portugal, 2013). http://purl.pt/24107/1/iPres2013_PDF/iPres2013-Proceedings.pdf

[8] Tarrant, D., and Van Der Knijff, J. 2012. Jpylyzer: Analysing jp2000 files with a community supported tool. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012). http://eprints.soton.ac.uk/341992/1/iPres2012.pdf

[9] Extensible Metadata Platform (XMP), http://www.adobe.com/products/xmp.html, Last Access: 04/16/2015.

[10] Apache Tika. https://tika.apache.org/. Last Access: 04/16/2015.