Date submitted: 07/06/2010

A data first approach to digital preservation: the SPAR project

Emmanuelle BERMES
Louise FAUDUET
and
Sébastien PEYRARD
Bibliothèque nationale de France
Paris, France

Meeting: **157. ICADS with Information Technology**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

In 2010, SPAR, the digital preservation repository of Bibliothèque nationale de France, has gone live. SPAR has been designed to ingest and preserve more than 1.5 petabytes of data from a variety of digital collections, including digitization, audiovisual and born-digital objects, as well as Web archives. The system is meant to be integrated into the day-to-day workflow of the library's activities, by allowing librarians to actively participate in the lifecycle management of their digital collections. BnF now benefits from the feedback of several years of work on the project, and can share the lessons learnt in developing this large-scale preservation system.

One of SPAR's major strengths is the way the system handles metadata. The data-first approach makes the system very flexible, as it is possible to change its behaviour by changing only the data, and not the software and processes. The system is scalable: it is possible to ingest new types of digital collections by just enhancing the core system if new requirements appear.

Finally, by creating a system which makes the data model a « lingua franca » between IT engineers, administrators, and librarians, SPAR has favoured the convergence of skills across the library. This includes the creation of the new position of « preservation expert » in the course of the project, at the intersection of the system and the data. These experts are responsible for the data model and the negotiation of SLAs with the librarians in charge of digital collections.

1. Context : BnF and the SPAR project

In 2006, the Bibliothèque nationale de France (BnF) launched a longstanding effort to build a preservation system[1] for its digital collections. Four years later, the system, called SPAR (Scalable Preservation and Archiving Repository) is finally stepping out of the design phases, and ingesting its first digital objects. Many lessons have been learnt in the course of the project, far beyond a strict digital preservation perspective.

One of the main challenges, when designing a system such as SPAR, is the risk of obsolescence of the system itself. When creating a software that is supposed to last several years, it is very difficult to guarantee that the evolutions both of the technological environment and of the digital objects to be preserved won't be so important that there will be a need to change the preservation system itself. Therefore, the main challenge is to be able to manage the digital collection, whatever evolutions may affect the information system, and in particular, its software components.

In order to shield ourselves from such an issue, from its early stages of conception, we designed SPAR as a modular system: in order to allow easier integration of new technology, each main function had to be able to be improved at its own pace. Thus the system was divided into modules following the OAIS functional model entities: Ingest, Data Management, Archival Storage, Access, Administration, and Preservation Planning, the last one to be developed at a later date. They form SPAR's "core".

Additional modules which do not have a direct equivalent in the OAIS functional model have been designed, such as a Rights Management module, which is not yet implemented, or Pre-Ingest modules for each specific set of homogeneous material. The Pre-Ingest phase is meant to harmonize the different digital documents into a SIP (Submission Information Package) that is SPAR-compliant and can be processed by the rest of the system in a generic way.

But this modular approach was not sufficient to ensure that the modules would actually be able to evolve independently from one another. They also needed to rely on a persistent data model, in order to be able to manage the data, even if the software changes. In a digital preservation perspective, letting the software handle the data in an opaque way is not acceptable. The OAIS model is mainly about making the processes that have an impact on data as transparent as possible. This is what we call the « data-first » approach [6].

This approach very much relies on the creation and maintenance of metadata. The system is fully self-describing: descriptions of the processes, agents (including software agents) and formats used in the system are ingested as information packages, themselves to be preserved. The behaviour of the system is managed according to policies that are agreed between librarians and administrators. The Data Management module holds all the information that is necessary to monitor the system, and to plan preservation actions.

Then again, the issue is not only the creation of metadata, but also its use. The traditional approach of metadata creation in the digital preservation community has been focused on

[1] By system, the authors mean the software part of the digital preservation framework. The hardware part, or infrastructure, has been set up at BnF starting in 2005 and is not described in this paper. See [1] for more information.

what type of metadata needed to be created [4], and which metadata standard would be most appropriate to store and exchange it [5]. Little has been said or written on how this data is actually to be used in preservation actions, or in the day-to-day management of the archive. The data-first approach we had in SPAR thus followed one major rule: if metadata is created, then it has to be used. This is what we will be explaining in this paper, by demonstrating how collection management is implemented in the system, and showing the benefits of the data-first approach from an organizational perspective.

2. Preserving objects, managing collections

2.1. The concept of "track"

One of the main concepts elaborated during the early stages of SPAR's design was the concept of "track". A track is a collection of objects that share the same requirements in relation with preservation. The following tracks were identified:

- preservation digitization
- audiovisual material
- automated legal deposit (Web archiving)
- acquisitions of digital contents
- third party archiving
- records management.

The criteria to identify a track were based on the homogeneity of the material, but also on high-level policies which defined the constraints and requirements associated with each track. For instance, a constraint on the legal deposit track is the necessity to accept all kinds of formats, because our legal obligations imply that we collect all the material that is produced, whatever its form and purpose. Another constraint for this track is the obligation to preserve this heritage forever; so no deletion of the original ingested data is possible in this track. On the contrary, for administrative documents, legal constraints force the deletion after a given period of time.

These examples show that the criteria that define the tracks are based on political and sometimes legal considerations, but at the same time, they have very direct and strong technical implications that the system must be capable of handling. Thus, the decision to group objects in tracks in order to manage their preservation policies is a major aspect of SPAR. The tracks are defined for a particular set of homogeneous digital material which require the same services from the system.

This is achieved through a set of formalized requirements that govern the relationship between the stakeholders of these particular sets of objects, and the administrators of the preservation system. They have to express the exact nature of their commitments to one another in a policy, in order to ensure the transfer of responsibility from the producer to the archive. This process guarantees a good knowledge of the risks attached to the digital objects, commits the producer to submit and ingest appropriate material, and ensures that the archive takes all the necessary actions in order to provide the preservation service.

2.2. The negotiation between producer and archive

From 2008 to 2010, we led the design of three tracks : preservation digitization, third-party storage (a subset of third-party archiving), and audiovisual material. We learnt very quickly that the “track” level, which we had envisioned to be the main level for managing the objects, was actually quite relevant from the stakeholder's point of view (producers and collection managers) but was unfit for technical purposes. We had to define smaller subsets of digital objects, that were homogeneous not only from a policy point of view, but from a technical point of view – in particular, we needed the objects to share the same formats, processes and storage requirements so that the system could apply global rules. This led us to define a smaller level of collection management, called the “channel”, where all objects share both the same policies and the same technical characteristics.

For each channel, the producer of the objects and the administrator of the archive negotiate three types of policies: one for ingest, one for preservation and one for dissemination. These policies are formalized procedures which help the producer to express his needs in quantifiable terms, which are then converted to formal rules to be used by the system.

At that level of description, it is possible to reach a point where the collection policies, as defined by the stakeholders, actually meet the requirements of the system from a technical point of view. Then it becomes possible to create machine-actionable data, that describes in a formal way the characteristics of a set of objects, that is, a channel, and the policies that apply to it. This data can be used by the system to determine actions and manage the digital objects. In SPAR, this is the role of the service level agreements (SLAs).

2.3. The Service Level Agreement

The service level agreement is a formal document that describes in an extensive way the processes, actors, content and strategies associated with a channel. It is accompanied by a “detailed technical notebook” describing precisely the structure of the objects to be preserved: in particular, the origin of the metadata, and the granularity of packages.

Each digital document is ingested into the SPAR preservation system as an Information package, as defined by the OAIS model, with a METS manifest as packaging information stored within each package. The overall implementation of METS for the digital collections ingested in the system is just one side of our data-first approach: we also need to describe the processes of the system and our choices in designing it, so that everything a preservation expert or digital curator needs to know about SPAR is documented within it.

Therefore, the policies (and thus the SLAs) are to be preserved within the system as well as the objects themselves, so that the system is completely self-described. In SPAR, each package belongs to a track, which can be viewed as a family of documents with similar intellectual and legal characteristics; each track has one channel per homogeneous technical characteristics[2]. Description of each channel and track is factorized in a dedicated

[2] For instance, the channel B of the Audiovisual track contains the product of the digitization of analog audio and video documents acquired through legal deposit, with well-described and easily manageable production formats; whereas the channel A of the same track concerns legal deposit of born-digital content (excluding documents harvested on the Web), which we are constrained to ingest "as is", with inevitably unknown or misused formats.

information package.

The channel packages contain the SLA in the form of 3 machine-actionable files:

- the ingest SLA (acceptable formats, volume, security levels…), which allows the validation of the producer's ingests, and formalizes the responsibilities of the archive for each category of format;
- the preservation SLA (retention time, assurance levels…), which defines where the archival information packages (AIP) are stored and how their lifecycle is managed;
- the access SLA (dissemination formats, time, availability…)[3].

The SLAs are written in XML, and define four types of requirements. Requirements at the channel level include the SLA's validity dates, the opening and closing hours of the service, or the maximum unavailability duration, for instance. There are also requirements on packages (minimum and maximum size of package, allowed and denied format types for the channel, AIP retention duration, and so on), on storage (number of copies, presence of encryption, etc.) and on processes, determining how the system's resources can be mobilized by the channel (minimum and maximum number of invocations of a process for a given period and so on).

All those requirements are entered into the Data Management module when a channel reference package is ingested. Then the other modules of SPAR can query this information in order to execute the tasks that require the checking of some of these parameters.

To see how this data is used in the daily workings of SPAR, and the Data Management module's role in them, we can take the "Ingest a SIP" use case as an example :

- Whenever the Ingest module receives notification of a new SIP, it is audited, and its METS manifest is validated using information from the channel package that has been ingested in the Data Management module: which users are authorized to submit packages in this channel, or what the METS profile for the SIPs of this channel is.

- The SIP's characteristics are checked against the channel's SLA, to check that ingest requirements such as the maximum size or the number of objects allowed in the package are met.

- The files are individually identified, characterized and validated. The result is compared with the list of formats accepted in the channel, listed in the SLAs. The behaviour of the system if the criteria of the SLAs are not met (rejection of the package or mere warning to administrators) is also specified in the SLA.

This use case shows that within SPAR, the concept of SLA is not only an abstract, organizational matter. It is embodied in a document in machine-actionable form, and the system actually uses this data to determine some of its more critical actions. This mechanism is fully part of what we call the data-first approach, because the metadata is not only informative, is it also used as settings for the system.

The main consequence of this is that the negotiation between producer and archive, as formalized by the SLA, provides a true guarantee to the collections' managers and stakeholders that the system will actually behave accordingly. For the administrators, this

[3] The Access SLA is not yet fully implemented in SPAR, since our Access module merely disseminates the AIP 'as is' for now.

facilitates the monitoring of the channel's workflow, and helps them determine which evolutions of the system rely on ingesting new SLA data, and which require modifications of the software. It increases trust in the system both from the librarians' and the administrators' point of view.

As for the persistence of the system, it is also improved, because the changes in the workflow of the producer (for instance the addition of a new production format, or the increase of the average package size) are translated into a change in the SLA, and not the code of the system. Software evolutions are limited, and the necessity of flexibility over time is delegated to the data.

3. Data and organizations : the benefits of the data-first approach

SLAs are just an example of the data being used by the system to monitor its critical functions. In addition to SLAs, there are other kinds of datasets that are ingested for system management purposes. As a whole, they form SPAR's reference information.

3.1. A reference track

SPAR being an OAIS-compliant system, every piece of preservation information has to be submitted and stored as an information package. To this end, it uses reference packages, of three different types: context, formats and agents.

- **Context** information relates to sets of objects: this includes the packages for tracks and channels, the latter containing the SLAs.
- We also give representation information about every **format** for which we have defined a preservation strategy and made a monitoring commitment. This can be standards such as TIFF 6.0, or BnF profiles restraining these formats, for instance uncompressed 24 bits TIFF in 300 dpi resolution.
- Finally, SPAR ingests reference information about **agents** achieving preservation operations, which can be human (administrator, preservation expert), software tools (identification, characterization and validation tools) and processes in SPAR (such as the ingest and package update process). In the future, we intend to use information packages to describe software environment in an emulation perspective.

Grouping information that is common to many digital objects is just one feature of reference packages. They also have maintenance enhancement advantages: updating this central information means not having to update every information package that relates to it.

As previously described with SLAs, these information packages allow us to set system parameters with machine actionable files. For instance, the system can check the conformance of image files with a specific profile of TIFF used at BnF (TIFF 6.0, 24bits, 300 dpi resolution, BnF watermarking, etc.) each time a package with files whose format is identified as TIFF is ingested. This way, data defines and configures processes, not the other way around. This enhances control of the system processes by users that are not IT

specialists. Last but not least, digital curators and preservation experts can find a sample file or the source code of a tool, with human readable description in each information package documentation about the format. Every aspect of the system functionalities that has an impact on librarianship is documented in SPAR.

As a matter of fact, the latest track born in SPAR, which we had not foreseen in the design phase, was actually a track dedicated to that kind of objects we include under the term of reference information. All this information about context, formats and agents is organized not only in information packages with a METS manifest, but also according to SPAR's data model in a track (reference track) with channels (context, format, agent), each channel coming with its own SLAs. In the end, we did create a new collection : a collection of reference information.

3.2. A global graph

As we demonstrated above, reference information in SPAR has been designed in order to facilitate the management of information within the system, by aggregating every piece of information that can be put together for a set of similar objects, thus avoiding redundancy. Moreover, this information is not only present in the system with informative purposes, but is used by the system to process ingested objects.

This could only be achieved through the use of a data model with flexible interlinking capacities, standard encoding and query mechanisms, and full independence from implementation choices. This is the reason why we chose RDF as the main standard to manipulate the data within SPAR.

The metadata repository is thus transformed from its submitted XML encoding into RDF when inserted into the Data Management module. The choice of RDF was made following a risk analysis based on the desired features of the main metadata repositories in SPAR [3]. Resource Description Framework has a very generic and versatile data model, where the information is expressed in triples, following the syntax subject/predicate/object. It came ahead in the analysis due to its very flexible query language, SPARQL, its good performances in mapping from the existing XML metadata, and its potential reversibility if we were to change the data format in the future.

RDF is natively adequate to manage data containing a lot of links, as is the case within SPAR due to the factorization of reference information in the reference packages. The use of XML is well fit for storing and exchanging files, and we find it very convenient to have a METS file within each package to gather all the metadata. But even if there are links within the METS files, and from one METS file to other packages, these links are lacking flexibility and are very demanding in terms of processing resources.

When the data is converted into RDF, every piece of information within a METS manifest becomes an independent triple, which can be understood and manipulated without having to parse the contextual XML information. In the end, we obtain a network of information where every object or piece of object is uniquely identified so that we can make assertions on it. Assertions that come from the object's METS manifest are seamlessly integrated with information that comes from reference packages, so that all the information in SPAR can be manipulated globally. The fact that information is stored in different packages in the beginning doesn't imply lower query performances or complex query models. The Data Management module contains the global graph of information that is needed to manage our

digital collections over time.

The advantages of RDF listed above are particularly valuable when it comes to data retrieval issues. Data is controlled, thus access is controlled : the same concepts and things always have the same name. Queries are precise because they go through controlled access points and structured data. And, contrary to what happens with relational databases technologies, it is not necessary to know the names of the categories of data in advance to formulate a query: they can be deduced from the way the data is structured, by successive queries.

Here are some examples of queries we can formulate about material from the digitized books and still images collections:

- Which packages have pages flagged as containing a table of contents, but no table of contents file in XML, which would allows dynamic navigation in the document? Answering this question helps plan retrospective creation of structured tables of contents.
- How many packages were ingested in SPAR the last month, and what are their number of files, the formats of those files, and the quality rate of their OCR? This conventional question shows that data also helps administrators monitor the system.
- Which packages in our digitization channel have invalid HTML table of contents files? Invalid HTML doesn't necessarily impede access to the document, but is certainly harder to preserve; such a query helps preservation experts plan invalid HTML files regeneration.

We see many advantages in the use of RDF to manage the data in our digital trusted repository, but we must also admit that there are remaining issues attached to adopting a relatively new technology. First, compared to other technologies, few software providers are available for RDF triple stores, and its implementation required a great amount of tuning and optimization. Its performances are also slower for the moment than those of traditional relational databases. Even though it may not be a foremost issue in a preservation perspective, quick response times give valuable comfort to digital curators. Moreover, tests conducted in 2008 showed that our triple store implementation reached its limits when the data volume nears 2 billion triples — although it should be noted that the performances of RDF technologies are improving steadily. Considering that the first channel of objects to be ingested in SPAR already includes an estimated quantity of 1 billion triples, we know that scalability will be an issue in the coming years.

3.3. An emerging function: the preservation expert

Another issue associated with the choice of RDF was the need for specific training, both on the IT and on the librarian sides.

On the IT side, Semantic Web technologies were previously unused at BnF, and require training, first for the digital preservation team, then for their collaborators. Day-to-day monitoring of the Data Management module is also more difficult, since there is little peer support or experience feedback yet.

On the librarian side, training issues are even greater, since SPAR is not only aimed at digital preservation experts, but also at producers of data-objects and collection curators. They have to understand SPAR's data model in order to express their information needs.

Ideally, everyone dealing with digital collections should be able to get the information they need directly from the Data Management module, which implies learning how to query it with SPARQL.

Moreover, the lack of well-established best practices in RDF modeling for digital preservation forced us to build SPAR's data model and the ontologies for expressing preservation metadata in RDF "on the fly", using common sense and professional experience in data modeling.

This "learning by doing" period saw the emergence of a new function among BnF staff, the function of preservation expert. Preservation experts have a librarian background but they acquired the technical knowledge necessary to understand the main functions of the preservation system, and in particular, its data model. They act as intermediates between the IT engineers and the digital collections' stakeholders, they help with the negotiation of the tracks and the specification of the channels, they analyze the data and define new data models, and last but not least, they manage an important part of the reference track, hence becoming curators of this new collection that has been created for the need of the system.

The fact that the system is designed in a data-first way has allowed preservation experts, once they had acquired the skills to create, design and query the data, to have more control over the way the system works and handles the collection. Hence in a long-term perspective, RDF has real organizational pros, as it allows the separation of technical/IT issues from data/librarian ones. As complex as RDF and SPARQL can be in the beginning, they give librarians a better control of their data, which also means, in a data-first approach, a better control of the system processes.

Ultimately, we hope that SPAR's data model, and its use of RDF technologies, will allow all BnF's staff dealing with digital collections preservation and curation to speak a common language that will adapt to different missions and different time constraints. Every person in interaction with the archive will have to refer to the same data model, using the same query language, whether they are planning long-term preservation actions such as migrations; have short-term decisions to make, requesting a new ocerization on certain documents for instance; or need the day's latest statistics. And eventually, all these users will have to define the necessary evolutions of the data model together.

4. Conclusion

If we consider SPAR today, we see a system where collections are managed in a balanced way by IT administrators and librarians, using as a "lingua franca" the global graph of information constituted by all the data ingested in the system, both metadata about the objects, and reference data about the collections. It allows us to monitor these collections in an effective way, by providing the collections' stakeholders with a visibility and transparency they scarcely had before.

But what about preservation? Today, preservation actions in SPAR are limited to quality improvement of legacy data before it is ingested in the system. The preservation planning module, which developments are planned for 2011, will allow us to plan, test and implement that kind of strategies for mass digital collections.

So, the major lesson learnt from SPAR is that digital preservation is not about preserving objects, it's about curating collections. If we want to curate a collection, we need to have

control of the data. We need to ensure that control by using the means available to us:

— organizational procedures, such as service level agreements, which allow for agreements between stakeholders on the actions that need to be undertaken;

— metadata, in the form that we first anticipated with the use of METS but also, in a much exacerbated way, in the form of reference information that can and must be factorized and used;

— standards, because they are a guarantee of persistence and reliability of data;

— technical infrastructure, but not as a black box: it is the system that depends on the data, and not the data that is designed to fit the processes;

— and finally, human resources, adequately trained and mobilized;

All of the above are part of an overall risk management approach, because in an environment where it is difficult to predict evolutions, every decision is taken while trying to maintain the risk level as low as possible.

The collection is mainly defined by the way its content is curated, and especially by the fact that the stakeholders of the collection are clearly identifiable. Policy statements are of little interest if they are not associated with an organizational vision of how the policies are going to be applied content-wise. As a matter of fact, within a specific track, curation decisions can only be taken according to the knowledge we have of the content, its intended audience, its specificity (rareness, fragility, etc.), exactly the way it was regarding traditional collections. A preservation decision is never only technical.

SPAR is not a secure vault where the digital collections will lie safe without the need to act on them. The system provides a reliable framework for bit-level preservation, and facilitates decision-making for preservation actions: the software doesn't undertake preservation actions by itself, but it makes them easier to design and implement. What we've built is a curation system, an instrument for managing a library collection in a trusted, durable way, starting with day-to-day activities such as collection quality monitoring and improvement. Will the digital collection be able to last decades or centuries? Today more than ever, this question appears less as a technical issue than as a true librarianship challenge.

REFERENCES

[1] Bermès, E. et al. "Digital preservation at the National Library of France: a technical and organizational overview", *World Library And Information Congress: 74th IFLA General Conference And Council*, 2008. Online at http://archive.ifla.org/IV/ifla74/papers/084-Bermes_Carbone_Ledoux_Lupovici-en.pdf [last consultation 2010-05-04].

[2] Bermès, E. and Fauduet, L. "The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France", *Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, 2009. Online at http://escholarship.org/uc/item/6bt4v3zs [last consultation 2010-04-20].

[3] Bermès, E. and Poupeau, G. "Semantic Web technologies for digital preservation: the SPAR project", *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008)*, 2008. Online at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_14.pdf [last consultation 2010-05-04].

[4] Farquhar, A. "Implementing Metadata that Guides Digital Preservation Services." *Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, 2009. Online at http://www.escholarship.org/uc/item/12p437bd [last consultation 2010-04-20].

[5] Guenther R., Wolfe R., "Integrating Metadata Standards to Support Long-Term Preservation of Digital Assets: Developing Best Practices for Expressing Preservation Metadata in a Container Format." *Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, 2009. Online at http://www.escholarship.org/uc/item/0s38n5w4 [last consultation 2010-04-20].

[6] Mazocchi, S.. "Data First vs. Structure First", *Stefano's Linotype*, July 28th, 2005. Online at http://www.betaversion.org/~stefano/linotype/news/93/ [last consultation 2010-06-01].