SPAR: From Design to Operations

Scalable Preservation and Archiving Repository (SPAR)
Agenda

- Main design choices
- Prepare for change
- Achievements and oversights
How to manage digital archives

• Build a central repository to diminish the diversity (media, formats, departments …)
• To take inspiration from good practices and standards

⇒ Key requirements:
  – OAIS compliance
  – modularity and distributivity
  – abstraction
  – use of well known formats and standards
  – use of open-source technical building blocks
## Missions of the library

### Missions:

- to build up the collections,
- to preserve and communicate them to the public,
- to produce a reference catalog,
- to cooperate with other institutions,
- to participate to research programs.

### Legal deposit:

- legal deposit since 1537 for printed materials
- 1648: engravings and maps
- 1793: musical scores
- 1925: photos
- 1938: phonograms
- 1975: videograms
- 1992: electronic documents
- 2006: Web legal deposit
SPAR
a generic repository solution at BnF
Decomposition in channels

• To deal with the variability and heterogeneity of the data, definition of **channels**
• build on the relation between the digital objects and the archival system, independently of any given organization:
  – Preservation digitization
  – Audiovisual material
  – Negotiated legal deposit (dark Web, regional press)
  – Automatic legal deposit (surface Web)
  – Administrative production
  – Deposit / Third party archiving
  – Acquisition / Donation
Planning

2009

2010

2011

2012

Preservation digitization

Third Party Archiving
Phase 1

Audiovisual

Third party
Phase 2

Web archiving

Evolutions in the core of SPAR

mise en production
mai 2010

Administrative production

{iBnF}
Storage abstraction

SPAR - Realization

AQS

AIP

Main storage

Backup storage

Multiple copies management

Refreshment of media

Media migration

Application viewpoint

Infrastructure viewpoint

SPAR - Infrastructure

SAS
Agenda

• Main design choices
• Prepare for change
• Achievements and oversights
Technological obsolescence

• Media migration:
  – Main site: from 9840C (40GB) to LTO5 (1.5TB)
  – Backup site: from T10000A (500GB) to T10000B (1TB)
  – 3 months operation:
    • migration of 208 000 AIP for 78TB
    • no downtime
    • in parallel, new ingestion of 26TB
Initial configuration

SU01

/Samfs

Main site

elemAIP01-1

elemAIP02-1

ONLINE

SU02

/Samfs

Backup site

elemAIP01-2

elemAIP02-2

ONLINE

{'BnF'}
Transitional storage unit

SU01

/samfs  Main site

elemAIP01-1

elemAIP01-3

(elemAIP01-2)

(elemAIP01-4)

{BnF}
Migration of the records

Migration made 1 full tape at a time, using the more dense ones as source.
Final configuration

SU01
- elemAIP01-3
- LTO5
- LTO5
- LTO5

ONLINE

SU02
- elemAIP02-3
- LTO5
- LTO5
- LTO5
- elemAIP01-4
- T10KB
- T10KB

ONLINE

/samfs
Main site

/samfs
Backup site

elemAIP02-4
- T10KB
- T10KB
- T10KB
Main benefit

From 2723 tapes (9840C)
Total weight : 764kg

To 70 tapes (LTO5)
Total weight : 18kg

Old capacity : 730TB
New capacity : 16 PB
Technological obsolescence

Triplestore version upgrade:
- 188 517 901 triples for 724 394 named graphs
- Just 3 days 😊
- but with a halt of the system 😞

RDF mapping modification:
- Need to reprocess 241 465 graphs
- Use of a planned program (rebuildMetadata)
<table>
<thead>
<tr>
<th>Voir le plan</th>
<th>CalculBilan001</th>
<th>Voir la définition</th>
<th>Voir les paquets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom du plan</td>
<td>CalculBilan001</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Type du plan</td>
<td>CalculBilan001</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Créé par</td>
<td>spar_admin1</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Décendé par</td>
<td>spar_admin1</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Nombre de paquets</td>
<td>49950</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Nombre de paquets OK</td>
<td>49950</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Nombre de paquets KO</td>
<td>0</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Temps moyen (HH:mm:ss)</td>
<td>00:00:02</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Date de début</td>
<td>2011/12/09 16:53:24</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Date de fin</td>
<td>2011/12/17 17:27:52</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Période</td>
<td>PT20S</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Priorité</td>
<td>low</td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
<tr>
<td>Raison de l'arrêt</td>
<td></td>
<td>Voir la définition</td>
<td>Voir les paquets</td>
</tr>
</tbody>
</table>
Agenda

- Main design choices
- Prepare for change
- Achievements and oversights
Main achievement

• Separation in channels:
  – Creation of a reference track with 3 channels
    • Every reference information is ingested in the system as a regular AIP
    • Formats, channels, agents (humans and tools)
  – Creation of a third-party track
    • Simple instantiation of a channel for each partner
    • Ability to ingest quickly every kind of material
Service Level Agreements

Which formats are allowed?

What's the maximum size of an SIP?

How many copies must I make, and where?

Do I need to log the dissemination requests?

BnF
Main oversight

- Main indicator: **number** of packages

```xml
<processRequirements>
  <comment>ING_1: Ingest process</comment>
  <processIdentifier type="ARK">
    ark:/12148/br2d27h
  </processIdentifier>
  <minByPeriod><!-- 5000 / day -->
    <numberOfEvents>5000</numberOfEvents>
    <period>PT24H</period>
  </minByPeriod>
</processRequirements>
```
Main oversight

• Another indicator is important: the size of the package
  – Packages are not uniform
  – Handling lot of small packages is very different from some big packages especially in terms of IO
Flow control

1 package every 9 is 10 times bigger

Process control:
3 packages at a time

Processing of small packages:
- 30 packages every 10

Processing with big packages in 10:
- 20 small packages
- 1 big package
Flow control

1 package every 10 is 10 times bigger

3 packages at a time

After 30, you end up having 3 big packages at the same time

=>

- requires 10 times the space to process
- forbids the small packages to enter
## Requirements on formats

<table>
<thead>
<tr>
<th>Kind</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>Stored</td>
</tr>
<tr>
<td></td>
<td>No technical information</td>
</tr>
<tr>
<td></td>
<td>Bit stream preservation</td>
</tr>
<tr>
<td>01</td>
<td>Identified</td>
</tr>
<tr>
<td></td>
<td>Identified format =&gt; associated mimetype</td>
</tr>
<tr>
<td></td>
<td>No preservation strategy planned by the institution</td>
</tr>
<tr>
<td>10</td>
<td>Known</td>
</tr>
<tr>
<td></td>
<td>Format identified, documented, with tools =&gt; associated schema of technical description</td>
</tr>
<tr>
<td></td>
<td>Preservation strategy planned by the institution</td>
</tr>
<tr>
<td>11</td>
<td>Managed</td>
</tr>
<tr>
<td></td>
<td>Identified format</td>
</tr>
<tr>
<td></td>
<td>Documentation and tools owned by the institution</td>
</tr>
<tr>
<td></td>
<td>Profile of use defined in the institution</td>
</tr>
</tbody>
</table>
# How to parameter such requirement

- **Allow/deny mechanism**
- **For each kind of format**

<table>
<thead>
<tr>
<th>Match</th>
<th>Allow, Deny result</th>
<th>Deny, Allow result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match Allow only</td>
<td>Accepted</td>
<td>Accepted</td>
</tr>
<tr>
<td>Match Deny only</td>
<td>Rejected</td>
<td>Rejected</td>
</tr>
<tr>
<td>No match</td>
<td>undefined</td>
<td>undefined</td>
</tr>
<tr>
<td>Match both Allow &amp; Deny</td>
<td>Rejected</td>
<td>Accepted</td>
</tr>
</tbody>
</table>
libmagic(3)
Example

```xml
<packageContent>
  <formatCategory type="identified" order="deny,allow">
    <formatList action="allow">
      <format type="MIMETYPE">image/jpeg</format>
    </formatList>
    <formatList action="deny">
      <format type="MIMETYPE">*/*</format>
    </formatList>
  </formatCategory>
  <formatCategory type="managed" order="deny,allow">
    <formatList action="deny"/>
    <formatList action="allow"/>
  </formatCategory>
</packageContent>
```
Conclusion

• Be always prepare for change at all levels (hardware, media, libraries, software)

• If something could occur, with the time going, it will occur

• Provision for valves to control the flow
Thanks for your attention

Questions?

Thomas Ledoux
thomas.ledoux_AT_bnf.fr