Preservation planning
at BnF

National Library of France
Bibliothèque nationale de France

March 12th, 2015
Agenda

• Context of digital assets at BnF
• People involvement
• Tests channel
• Conclusion
The National Library of France

An autonomous public establishment

2200 agents and dozens of professions, a budget of aprox. 230 M€

Local, national and international missions

About 1M readers per year, 300 000 visitors to the exhibitions

Large collections to manage

- More than 14M books
- More than 30M posters and photos
- More than 250,000 manuscripts
- More than 1M audiovisual material
- “French” Web legal deposit
Variety of the collections
Digital material

Two paths of acquisition:

1. Digitization
   - mean for minimizing the risks on media, devices and formats
   - offering a digital surrogate limits the access to the original

2. Born-digital document
   - audiovisual legal deposit
   - web legal deposit
   - the document belongs to the heritage collections, to be preserved and given access
Digital preservation at BnF: SPAR

Production applications
- Preservation digitization
- Audiovisual
- WEB Archiving

Dissemination applications
- Gallica (digital library)
- Public workstations
- wayback

SPAR - Infrastructure
- SPAR - Realization
- Storage Abstraction Service (SAS)
- Ingest
  - Data management
  - Administration
  - Storage
  - Access
  - Preservation planning

03/12/2015
SPAR - PASIG 2015 - San Diego
SPAR
a generic repository solution at BnF

Preservation digitization

Audiovisual collections

And so on

Preservation planning
Administration

Data management

Access

Ingest

Storage abstraction service

Storage

Optional Transformation step

SIP

DIP

Infrastructure

03/12/2015

SPAR - PASIG 2015 - San Diego

Slide 7
Decomposition in tracks

- To deal with the variability and heterogeneity of the data, definition of **tracks**
- build on the relation between the digital objects and the archival system, independently of any given organization:
  - Preservation digitization
  - Audiovisual legal deposit
  - Negotiated legal deposit (e-books, large posters, …)
  - Automatic legal deposit (surface Web)
  - Administrative production
  - Third party archiving
  - Acquisition / Donation
Agenda

• Context of digital assets at BnF
• People involvement
• Tests channel
• Conclusion
Tracks currently in operation  
(since may 2010)

- Preservation digitization: in doors and out doors
  - 1.7 million AIPs for 1.5 PB
- Audiovisual material
  - 97 816 AIPs for 83 TB
- Large posters legal deposit
  - 77 AIPs for 317 GB
- Web archiving
  - 1.8 million AIPs for 195 TB
- Third party archiving: 3 contracts
  - 65 000 AIPs for 7 TB
- Reference channels (SLAs, formats, tools, tests)
Evolution in the organization

The library is building a new organization to interact with the Archive:

• census of preservation experts (expertise everywhere in the library)

• each track has a designated manager from the business units, to negotiate the SLAs and define the priorities

• the IT department is responsible for the data in the “digital shelves”

Most importantly, preservation is now taken into account very early in the new projects
People involved in preservation planning

- Track Manager
- Administrator
- Preservation Expert
- Developer
- Collection Manager
- Risks & Emergency Plan
The Service Level Agreements (SLAs)

- Which formats are allowed?
- What is the maximum size of a package?
- How many copies are needed, in what kind of media?
- Do we need to log each access?

Storage abstraction service
Agenda

- Context of digital assets at BnF
- People involvement
- Tests channel
- Conclusion
PRESERVATION PLAN
(transform to pdf)

TEST CAMPAIGN
File conversion
Under Windows
Microsoft & Open Office formats
to
pdf format
(tool: pdf creator)

TEST CAMPAIGN
File conversion
Under Linux
Microsoft & Open Office formats
to
Open Office format
(tool: Libre Office)

TEST CAMPAIGN
File conversion
Under Linux
Microsoft & Open Office formats
to
pdf format
(tool: Libre Office)
TEST CAMPAIGN: Transformation of files under Linux MS & OO to pdf format (tool: Libre Office)

Test data package (Initial data) → Characterization tool Tika 1.5 → Test Metadata package (Characterization of initial data)

Transformation tool Office to pdf

Test data package (Transformed data) → Characterization tool Tika 1.5 → Test Metadata package (Characterization of transformed data)
Description of tika

Identifiant de production *: info.bnf/spar/agent/tika
Catégorie: Agent

Nom *: Outil Tika
Alias *: tika

Type:  
- human
- module specification
- softwareAgent
- process

Sous-type:
- identification tool
- characterization tool
- validation tool
- migrationTool

Module

Ark du paquet de Module
Rechercher...

Source

URL référence: http://tika.apache.org/
Date: 2013-09-02
URL download: http://tika.apache.org/download.html

Descriptions *

Outil d'identification et de caractérisation de format [fr]

(+) Ajouter
# Environment for tika

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identifiant de production</strong></td>
<td>info:bnf/spar/agent/tika</td>
</tr>
<tr>
<td><strong>Catégorie</strong></td>
<td>Agent</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>1.6</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2014-10-03</td>
</tr>
<tr>
<td><strong>Ligne de commande</strong></td>
<td>java -jar tika-app-1.6.jar -xmp</td>
</tr>
<tr>
<td><strong>Point d'entrée</strong></td>
<td>tika-app-1.6.jar</td>
</tr>
</tbody>
</table>

## Characteristics

- **known to work**
- **run**

## Dependency

- **Nom**
  - java

## Software

- **Nom**
  - JRE
- **Version** >=1.5
- **Type**
  - java runtime environment

## Additional Information

- **Java Runtime Environment au moins 1.5**
Channel reference package

• 3 SLAs: Ingest, Preservation, Access
• Formalize in XML the ways of managing the packages
• Those 3 SLAs are recorded in a reference package that describes the channel

Mets.xml
Contract.pdf
SLA-I.xml, SLA-P.xml, SLA-A.xml
Agenda

• Context of digital assets at BnF
• People involvement
• Tests channel

• Conclusion
Conclusions

• Preservation planning is a human activity and needs an organization around it
• Need to keep track of each decision made for the long term
• Need to enforce the decision in a machine actionable way
Thanks for your attention

Questions?

thomas.ledoux_at_bnf.fr