Lessons from the implementation
Missions and context

From theory to practice
  → the model
  → the expectations
  → handling “bad” data
  → lessons

Management through risk analysis
  → Focus shift
  → Dealing with human skills
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
Digital archiving at BnF

Production applications

Preservation digitization
Record management
WEB Archiving

Dissemination applications

Gestion des droits
Préservation
Gestion des données
Administration
Stockage
Accès

Service d'Abstraction de Stockage (SAS)

wayback

24/06/2009 Distributed Archiving & Preservation System (SPAR)
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
Agenda

- Missions and context
- From theory to practice
  - the model
  - the expectations
  - handling “bad” data
  - lessons
- Management through risk analysis
  - Focus shift
  - Dealing with human skills
Each channel is formally defined by a reference package:
- description of the Service Level Agreement (SLA)
- includes human readable definition
- machine actionable parameters
- links to accepted formats at various levels of commitment (stored, identified, known, managed)
- links to the used tools
<sla:serviceLevelAgreement>
  <sla:header>
    <sla:channelIdentifier>FIL_NUM_CONS_A</sla:channelIdentifier>
    <sla:type>info:bnf/spar/context/channel#ingest</sla:type>
  </sla:header>
  <sla:packageAttribute>
    <sla:minSize unit="kilobyte">10</sla:minSize>
    <sla:maxSize unit="gigabyte">40</sla:maxSize>
    <sla:maxNumberOfFiles>3000</sla:maxNumberOfFiles>
  </sla:packageAttribute>
  <sla:packageContent>
    <sla:formatCategory type="info:bnf/spar/representation#storedFormat"
                        order="deny,allow">
      <sla:formatList action="deny"><format>*</format></sla:formatList>
    </sla:formatCategory>
    <sla:formatCategory type="info:bnf/spar/representation#managedFormat"
                        order="deny,allow">
      <sla:formatList action="allow">
        <format type="ark">ark:/12148/fTIFF_6_0w</format>
      </sla:formatList>
      <sla:formatList action="deny"><format>*</format></sla:formatList>
    </sla:formatCategory>
  </sla:packageContent>
</sla:serviceLevelAgreement>
Expectations

■ For the digitized collection,
  ➔ all formats are managed

■ BUT we deal with 10 years of data creation
  ➔ historical choices
  ➔ changes in requirements
  ➔ changes in scope
  ➔ learning mechanisms
Testing the formal method, we discover “bad” HTML files => not valid against the W3C validator

BUT these files have been
- produced for 2 years
- displayed for 8 years

Possible strategies:
- correct the data before ingestion
- accept “bad” data and plan for correction
Previous experience: SGML files

- Production made on SGML format (special TEI profile)
- Uncontrolled conversion made in XML (the current format at the time)
- At ingestion time, no way to display correctly the files

- Need for *archeological* search, to come up with:
  - The original XML files
  - The original specifications (SGML DTD)
  - The expected XML files
  - The needed tools (SGML parsers)
  - The required skills
The idea is to accept the files in a control way

- Lax the requirements on the channel (some identified formats)
- Use the ingest mechanisms to characterize the files as much as possible
- Formally determine the set of “bad” data
- Plan for *curation* with a well defined and documented migration plan

Benefits

- the history of transformation is preserved, the decision are traced and the tools used are known
Lifecycle of an Archival Package (AIP)

- Each version/edition has its own internal identifier
- The persistent identifier is unique, in ark: format

V0 → V1 → V2 → V3 → V3’ → V4 → V4''

Destructive operation on a data-object: update or deletion

1/ Update of metadata only (PDI, IR, etc.): addition, update, deletion
2/ Addition of data-object

Digital original

ark:/12148/…
Some lessons

- **Lesson 1:**
  - never think your data is perfect

- **Lesson 2:**
  - the model works !!!

- **Lesson 3:**
  - migration is part of the preservation process
  - keep the original data and trace the operations made
Instructions to use the original tool

**Instruction pour se servir des disques reproduisant les ondes sonores.**

On voit sur ces disques des sillons représentant des ondes sonores produites par la voix humaine.

Pour obtenir la reproduction des sons émis et ainsi enregistrés, il suffit de poser le disque sur le plateau d'une machine parlante et de faire tourner à ce plateau environ 70 révolutions à la minute.

La machine parlante se compose des organes suivants:

1. boîte; 2. diaphragme; 3. membrane; 4. porte-aiguille; 5. aiguille; 6. plateau; 7. disque; 8. bras acoustique; 9. pavillon; 10. manivelle.

Le mécanisme moteur est de préférence un moteur à ressort enfermé dans la boîte 1 formant l'embase au socle de l'appareil.

L'organe répétiteur du son est constitué par une boîte de résonance ou diaphragme 2, affectant la forme d'un cylindre court, en métal ou autre matière, comportant sur l'un des côtés une membrane 3 en ivoire ou autre substance. Au centre de cette membrane s'adapte une...
Agenda

- Missions and context
- From theory to practice
  - the model
  - the expectations
  - handling “bad” data
  - lessons
- Management through risk analysis
  - Focus shift
  - Dealing with human skills
Risk analysis

- Risk management is at the core of preservation strategy
- Necessity for reassessing the risks on a regular basis
- Inclusion of the preservation system in the whole organization:
  - sustainability of the economics
  - add value to the service
Handling the transition

- Building a smooth transition in two phases
  - ingesting and auditing
  - direct access

- BUT dissemination has became critical in our ecosystem:
  - access through Gallica
  - cooperation with Europeana
  - experimentation with publishers

- Risk to high to modify access at the moment
Focus shift to preservation repository

- The SPAR project focus shifts to
  - ingesting data
  - storing it safely
  - and managing it
- It acts as a digital repository for preservation
Dealing with human skills

- Need for enhanced skills in the library
  - preservation expert: digital archeologist, model and formats
  - digital curator: management of digital collections
  - computer scientists: advise on tools
  - administrator: dealing with the data deluge
  - manager: understanding of digital issues, endorsement of channels
Implementation

Pre Ingest

Ingest

Preservation

Administration

Rights management

Access

Data management

Storage

SAS

JHOVE

magicmimetypeidentifier

Java Advanced Imaging

Struts$^2$

OpenLDAP

VIRTUOSO

PostgreSQL

Spring

LogBack

JBoss

Jersey

SPAR
Thank you for your attention

Questions?

More information: http://bibnum.bnf.fr/spar

Thomas Ledoux
thomas.ledoux_AT_bnf.fr