Bibliothèque nationale de France (BnF)

Distributed Archiving & Preservation System
Système de Préservation et d’Archivage Réparti (SPAR)

SPAR Architecture Developments

Sun-PASIG

07/10/2009
Agenda

- Missions and context
- Overall architecture
- Management of metadata
- New channels:
  - Third party storage
  - Audiovisual material
- Organizational shift
- Conclusion
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 in 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

Dissimination applications

Preservation digitization

Audiovisual

WEB Archiving

 wayback
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
- Overall architecture
- Management of metadata
- New channels:
  - Third party storage
  - Audiovisual material
- Organizational shift
- Conclusion
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
Backup site

Main site

07/10/2009 Distributed Archiving & Preservation System (SPAR)
BnF estimates

- third party archive
- reproduction
- still images
- mass digitization
- newspaper
- internal digitization
- B&W digitization
- web archive
- video conversion
- audio conversion

Estimates
Missions and context
Overall architecture
Management of metadata
New channels:
  → Third party storage
  → Audiovisual material
Organizational shift
Conclusion
In order to query the metadata, those have to be extracted from the manifest.

The METS files contains also reference metadata in case of problem, metadata can be retrieved from the AIP.
Data management use cases

DATA MANAGEMENT

DM_1: Search MD

DM_1.1: SimpleSearch  DM_1.2: DeepSearch

<<include>>

DM_2: Store MD

DM_3: Retrieve MD

DM_4: Monitor DM

DM_5: Delete MD

DM_6: Reference Document Access
Use of TripleStores

- Data model: mapping from METS to RDF
- Modularity and flexibility:
  - Information is represented as simple sentences
  - Each "sentence" is queriable in the RDF store
  - Request can be built as needed

- Ontologies
  - RDF triple build with respect to ontologies
  - New ontologies based on the OAIS model (structure, representation, provenance, context, …)
  - Gives independence in regards to the manifest representation
Data repository

Metadata repository
- Complete
- Selection
- Reference

Reference data

rebuild procedure
Limitations

- Triplestore able to deal reliably with few billions of triples
- The digitization channel will cover about 1 million package with a mean of 100 objects and 200 files.
- If we have 20 triples per file, this means 4 billion triples just for the information on the files !!!
- In order to cope with this issue, we aggregate the file information at the group level:
  - when we search, we only need to look for what is in the package
  - when we need to manipulate the content of a package, we retrieve all the content including the manifest
Agenda

- Missions and context
- Overall architecture
- Management of metadata
- New channels:
  - Third party storage
  - Audiovisual material
- Organizational shift
- Conclusion
Challenges of new channels

- Introducing new channels means adjusting the model
- Each channel should integrate smoothly in the system and bring its own features
- Third party storage
  - Accept any file: even not recommended ones
    - Turn off strong characterization
  - Deal with many producers and different ways of exchange
    - Automate the SLA writing
    - Try to extract at least descriptive metadata
    - Need monitoring for each channel
Challenges of new channels

- Audiovisual
  - complex hierarchical structure
  - Multimedia
  - multivolume
  - Same magazine

Each media produces its own digital artefact: different workflow of digitization
Challenges of new channels

- Audiovisual
  - variety of formats
  - Contradiction in the mandate
    - legal deposit: everything must be accepted
    - dissemination: need for known formats
  - Introduction of 3 related channels
    - channel FIL_AUD_A: all legal deposit material
    - channel FIL_AUD_B: material controlled by the library (digitization, extraction)
    - channel FIL_AUD_C: digital surrogate for dissemination (may need extensive work)
  - need for adequate validation tools
  - need for new technical metadata schemas
Agenda

- Missions and context
- Overall architecture
- Management of metadata
- New channels:
  - Third party storage
  - Audiovisual material
- Organizational shift
Shift in focus: initial view

Document digitization

Ingest

Dissemination

SPAR

Rights

Metadata
In this first phase, the archive packages in SPAR won’t be disseminated in an intense way => focus on ingesting and auditing.
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
  - storage expert: dealing with the storage infrastructure
  - manager: understanding of digital issues, endorsement of channels
Agenda

- Missions and context
- Overall architecture
- Management of metadata
- New channels:
  - Third party storage
  - Audiovisual material
- Organizational shift
- Conclusion
Overall ingestion of a package
Thank you for your attention

Questions ?

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

Thomas Ledoux
thomas.ledoux_AT_bnf.fr