Information model
and
technical building blocks

Sun-PASIG San Francisco
Agenda

- Digital archiving at BnF
- Information model
  - structure of a package
  - reference packages
- Building blocks
- Conclusion
Digital archiving at BnF

Production applications
- preservation digitization
- digital assets
- web crawler

Dissemination applications
- Gallica
- images
- wayback

Ingest
- Metadata
- Administration
- Storage

SIP
- AIP
- DIP

PRESERVATION

STORAGE INFRASTRUCTURE
Decomposition in channels

- Build on the relation between the digital objects and the archival system, independently of any given organization:
  - Preservation digitization
  - Reproduction digitization
  - Automatic legal deposit (surface Web)
  - Negotiated legal deposit (dark Web, regional press)
  - Administrative production
  - Deposit / Third party archiving
  - Acquisition / Donation
Agenda

- Digital archiving at BnF
- Information model
  - structure of a package
  - reference packages
- Building blocks
- Conclusion
Information model

- Packaging made through METS:
  - Descriptive metadata: qualified Dublin Core
  - Preservation metadata: PREMIS
  - Rights information: ODRL
  - Technical metadata: dependant on the channel
    - MIX, textMD, ...

- 4 levels of granularity:
  - set: intellectual grouping (collection, periodical, ...)
  - group: digital object (monograph, series of images, ...)
  - object: digital element (a page, an image, a track, ...)
  - file: data-object (digital file or bit-stream)
Archival Information Package

- **Package Description**
- **Archival Information Package**
- **Packaging Information**
- **Content Information**
- **Preservation Description Information**
- **Data Object**
- **Representation Information**
- **Reference Information**
- **Fixity Information**
- **Physical Object**
- **Digital Object**
- **Structure Information**
- **Semantic Information**
- **Provenance Information**
- **Context Information**

Interactions:
- **derived from**
- **described by**
- **identifies**
- **interpreted using**
- **adds meaning**
Reference packages

- **Strongly linked metadata**
  - directly attached to the files that constitute the package
  - describes the formats, the audit trail, fixity and identity information, ...

- **Weakly linked metadata**
  - reference information that can be shared by multiple packages
  - specifically
    - channel description (SLA)
    - information representation (formats)
    - process description (tools)
    - transformation description
    - migration plan
Sample strongly/weakly linked

```
<dmdSec ID="DMD.0001">
  <mdRef LOCTYPE="ARK"
xlink:type="simple"
xlink:href="ark:/12148/bc343631419"
MDTYPE="DC"
MIMETYPE="text/xml"/>
</dmdSec>

<dmdSec ID="DMD.0002">
  <mdWrap MIMETYPE="text/xml" MDTYPE="DC">
    <xmlData>
      <dc:dc>
        <dc:title>La Croix</dc:title>
        <dc:date>1883/12/12</dc:date>
        <dc:type>PERIODIQUE</dc:type>
      </dc:dc>
    </xmlData>
  </mdWrap>
</dmdSec>
```
Channel description

- 4 objects:
  - general description
  - ingest service level agreement
    - SIP structure, categories of formats
    - volume, availability schedule
  - preservation service level agreement
    - length of preservation, level of preservation, number of copies
  - access service level agreement
    - DIP structure
    - volume, availability schedule

- each time: two representations
  - literal description: text of the Service Level Agreement
  - formal description: XML file that gives the exact parameters
Channel description (sample)

```xml
<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">42</sla:minSize>
    <sla:maxSize unit="gigabyte">5</sla:maxSize>
    <sla:maxNumberOfFiles>32</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:formatCategory>
  </sla:packageContent>
</sla:serviceLevelAgreement>
```
## Four categories of formats

<table>
<thead>
<tr>
<th>Code</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>Stored</td>
<td>No technical information&lt;br&gt;Bit-stream preservation level</td>
</tr>
<tr>
<td>01</td>
<td>Identified</td>
<td>Format identified&lt;br&gt;No preservation plan</td>
</tr>
<tr>
<td>10</td>
<td>Known</td>
<td>Format identified, documented, with tools and under monitoring by BnF experts</td>
</tr>
<tr>
<td>11</td>
<td>Managed</td>
<td>Format identified, documented, with tools&lt;br&gt;BnF has commitment on this format</td>
</tr>
</tbody>
</table>
1 object with two representations

- literal description:
  - text of the standard (known format, eg TIFF v6) or
  - text of the specific use (managed format, eg BnF specific BW G4 TIFF)

- formal description:
  - specific XML file that gives the required information
  - identification properties
  - schema of characterization
  - characterization properties
  - reference tools and expected behavior

- to be replaced by the GDFR standard …
Format description (sample)

```xml
<format>
  <name>TIFF BnF Noir et blanc</name>
  <category>info:bnf/spar/representation#managedFormat</category>
  <identification>
    <property type="magicname"><value>TIFF</value></property>
    <property type="mimetype"><value>image/tiff</value></property>
  </identification>
  <characterization>
    <schema type="characterization">
      <name>MIX</name>
      <namespacePrefix>mix</namespacePrefix>
      <namespaceURI>http://www.loc.gov/mix/v10</namespaceURI>
    </schema>
    <properties>
      <property id="compressionScheme">
        <name>CompressionScheme</name><value>4</value>
        <xpath>//mix:Compression/mix:compressionScheme</xpath>
      </property>
      ...
    </properties>
    <referenceTools>
      <tool type="identification">
        <identity>ark:/12148/bfile1</identity>
        <outcome propertyRef="magicname" use="contains" />
      </tool>
    </referenceTools>
  </characterization>
</format>
```
1 object with two representations

- literal description:
  - text introducing the tool or process used

- formal description:
  - XML file that gives the required information
  - based on the Environment PREMIS tag
  - exact version of the tool
  - execution platform
Agenda

- Digital archiving at BnF
- Data model
  - structure of a package
  - reference packages
- Building blocks
- Conclusion
Modularity of SPAR
Main functionalities

- Import of SIPs
  - from all the channels
  - from reference description

- Management of updates
  - new files in a package
  - updates of certain files

- Controls
  - Integrity
  - File formats
  - METS (.xml)

- Generation of the unique identifier and making of the AIP

- Dialog with the others modules
  - Data management
  - Storage

Implementation

- Building blocks:
  - JHOVE
  - axis
  - NOID
  - Struts
  - schematron

- Specific developments:
  - METS manifest enrichment
  - Calls to DM and STO
  - Transaction (ACID) management

27/05/2008
Distributed Archiving & Preservation System (SPAR)
Main functionalities

- Storing an AIP
  - SLA search
  - Dialog with SAL
  - Store the AIP and send acknowledgment

- Audit of an AIP
  - Verification of the package itself
  - Integrity checks on the files

- Retrieval of an AIP
  - Dialog with Access module
  - Asynchronous retrieval
  - Provide a copy of the AIP

Implementation: Storage and Storage Abstraction Service (SAS)

- Building blocks:

- Specific developments:
  - Transaction (ACID) management

Storage module
Data management module

» Main functionalities

- Search on metadata
  - Simple and synchronous search
  - Deep search (any metadata)

- Metadata storing
  - Index update

- Retrieval of metadata
  - Dialog with Access module
  - Retrieval of the metadata

» Implementation

- Building blocks:
  - mulgara (RDF)
  - Powered by VIRTUOSO

- Specific developments:
  - Transaction (ACID) management
Rights management module

Main functionalities

- Harvest rights information from SOLON (BnF specific rights management system)
  - Rights metadata
  - Decision trees and agreements
  - Updates

- Generate a license
  - Retrieve MD for a DIP
  - Traverse the tree or apply the agreement
  - Generate license (ODRL)

Implementation

- Building blocks:
  - woodstox
  - Proai

- Specific developments:
  - Transaction (ACID) management
Main functionalities

- DIP export
  - Synchronous
    - Management of DIP cache
  - Asynchronous
    - Build the DIP with the assistance of the other modules and send it
- Search in the metadata
  - Result lists: persistent identifier of packages
- Harvest metadata
  - Retrieve the MD of a package

Implementation

- Building blocks:
  - woodstock
  - Struts²

- Specific developments:
  - Call of DM and storage modules
  - Transaction (ACID) management
Main functionalities

- Management and federation of identities
- Description of reference description
  - SLA and channels descriptions
  - Process descriptions
- Planning
  - Access plans
  - Storage audit plans
  - MD reconstruction plans
  - Migration plans
- Plan monitoring
- System monitoring
- Administrative data export
- Accounting data export

Implementation

- Building blocks:
  - Struts
  - Log4J
  - OpenLDAP
  - Quartz
- Specific developments:
  - Journal of actions
  - …
Main functionalities

- Description
  - Representation information
  - Transformation
  - Migration plans
- Harvest of representation information from a format registry
- Monitor a migration plan

Implementation

- Building blocks:
  - Struts²
  - Proal
  - OAIHarvester2
- Specific developments:
  - Journal of actions
  - ...
General overview of the building blocks
Conclusion

- Goal for the archived objects
  - definition of an open model
  - completeness of the description
  - self-supporting package

- Ways of dealing with the permanency
  - modularity
  - abstraction
  - use of well known formats and standards
  - use of Open Source technical building blocks
Thank you for your attention

Questions?

Thomas Ledoux
thomas.ledoux_AT_bnf.fr