Linked data and preservation metadata

A work in progress at BnF
The basis: persistent ids

• Global identification of
  – our digital assets
  – descriptions of our digital assets

• Need for (very) long term persistence
Identifying our assets: ARK

Opaque
Globally unique

Thought for long-term

Strong commitment from BnF to ensure persistent access

http://gallica.bnf.fr/ark:/12148/bpt6k107371t/f134.thumbnail

• a digital asset

http://catalogue.bnf.fr/ark:/12148/cb13894249p

• an authority record

http://catalogue.bnf.fr/ark:/12148/cb37982960c

• a bibliographic record
Preserving our assets
the SPAR repository

Pre Ingest

Digitized books

Digitized audiovisual documents

web archiving

Pre Ingest

Pre Ingest

Pre Ingest

METS

SIP

METS

Ingest

Preservation Planning (not yet developed)

Administration

Data Management

Storage

Storage Abstraction Services

Access (AIP copy)
Identification in SPAR

- **All** information packages in SPAR must an ARK id
  - Pre-existing: SPAR retrieves the ARK identifier
  - Otherwise: SPAR mints it
Factorizing information the "reference" packages

- Some information is used by a lot of assets
  - **Policies** and service level agreements
  - **Formats** that we know to handle
  - Preservation **tools** that we use

- Needs to be factorized, citable, and preserved

**OAIS part of the problem:**
Must come as an information package

**Citation part of the problem:**
Needs unambiguous & persistent identifiers

An example:

```xml
<premis:object xsi:type="file"> […]

<premis:format>
  <premis:formatDesignation>
    <premis:formatName>image/tiff</premis:formatName>
    <premis:formatVersion>6.0</premis:formatVersion>
  </premis:formatDesignation>
  <premis:formatRegistry>
    <premis:formatRegistryName>BnF SPAR</premis:formatRegistryName>
    <premis:formatRegistryKey>ark:/12148/bd2d2xn</premis:formatRegistryKey>
  </premis:formatRegistry>
</premis:format>

[...] </premis:object>
```
Linked data in SPAR

- In all AIPs
  - a METS wrapper
  - with PREMIS inside

- Our data management
  - a big linked data pool
  - an « internal interoperability tool »
Why RDF?

• Standardized
  – Stability of specifications (W3C recommendations)
  – Made to work together (RDF, RDFS, OWL, SPARQL…)
  – Independent from any implementation

• Based on globally unique identifiers (URIs)
  – All information must be globally unambiguous

• Great power and querying flexibility
  – Based on the conceptual data model (the physical implementation can change)
  – Lowers the barrier for non IT guys
  – Seamless data continuum from one package to another

→ Semantic web technologies chosen for persistency reasons
Identification in RDF

• We need URIs for an information package

• We need URIs at a lower level
  – versions of a package
  – parts of a package
  – events concerning all or part of the package

• We need URIs for classes and properties
  – Preferably not opaque

→ Need for a URI policy in SPAR
→ ARK does not solve all problems
URIs for... AIPs

• The URI for a package: the persistent part of the ARK identifier
  – ark:/12148/bpt6k102002g
  – Not a IANA-registered URI…
  – But a persistent one, that complies with the URI syntax
URIs for... AIP versioning

DO

version0

Destructive operation: update, deletion

version1

DO

version2

Reversible operation:
1/ Metadata update (PDI or RI): addition, correction, deletion
2/ Addition of data objects

version2

release1

version2

release2

Digital master (never destroyed)

ark:/12148/bpt6k102002g.version0.release0  ark:/12148/bpt6k102002g.version1.release0  ark:/12148/bpt6k102002g.version2.release0

ark:/12148/bpt6k102002g.version2.release1  ark:/12148/bpt6k102002g.version2.release2
URIs for... parts of an AIP

- A concrete AIP
  ark:/12148/bpt6k102002g.version0.release0

- An abstract sub-object in it (e.g. a page)
  ark:/12148/bpt6k102002g/f1.version0.release0

- A representation of this object
  - ark:/12148/bpt6k102002g/f1/master.version0.release0 (master image file)
  - ark:/12148/bpt6k102002g/f1/ocr.version0.release0 (OCR file)
To describe our packages in RDF, we need another URI scheme that allows significant identifiers.

- **Choice**: info:URI
  - info:bnf/spar/<ontology>\#dictionaryElement
  - info:bnf/spar/<ontology>/instance

- **An example of a class**: info:bnf/spar/provenance #ingestCompletion

- **An example of an instance**: info:bnf/spar/representation/tiff_6_0
Designing the ontologies

• One ontology per information type
  – OAIS structure, provenance, reference, fixity, representation and context
  – agent, taken from PREMIS
  – specific technical information: textMD, MIX…

• Re-use classes and properties whenever relevant
An ontology: provenance
From METS-PREMIS to RDF

<ark:/12148/bpt6k206840w.version0.release0>
  rdf:type sparstructure:group>
  dc:title "Impressions de Sicile" ;
  dc:creator "Volkonskaïa, Mariä" ;
  spar/provenance:hasEvent
  <info:bnf/spar/provenance/f4ca87f0-8453-11df-8668-00144f68e1cc>.
  <info:bnf/spar/provenance/f4ca87f0-8453-11df-8668-00144f68e1cc>
    rdf:type spar/provenance:ingestCompletion ;
    dc:date "2010-06-30" ;
    spar/provenance:hasPerformer
  <ark:/12148/br2d27h/act09>.

WHERE {
?package rdf:type sparstructure:group ;
  dc:title "Impressions de Sicile" ;
  dc:creator ?creator ;
  spar/provenance:hasEvent
  <info:bnf/spar/provenance/f4ca87f0-8453-11df-8668-00144f68e1cc>.
  <info:bnf/spar/provenance/f4ca87f0-8453-11df-8668-00144f68e1cc>
    rdf:type spar/provenance:ingestCompletion ;
    dc:date "2010-06-30" ;
    spar/provenance:hasPerformer
  <ark:/12148/br2d27h/act09>.

... and the answer is...

<table>
<thead>
<tr>
<th>package</th>
<th>digitization</th>
<th>performer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ark:12148/bpt6k206840w.version0.release0</td>
<td>Volkonskaïa, Mariä</td>
<td>28/07/2010</td>
</tr>
</tbody>
</table>
Querying the data...

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```sparql
SELECT DISTINCT ?tool ?name ?toolType WHERE {
?fileGroup a sparstructure:fileGroup;
    sparprovenance:hasEvent ?event.
?event a sparprovenance:fileProcessing;
    sparprovenance:hasPerformer ?tool.
?tool a ?toolType;
    foaf:name ?name
}
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML

Execution timeout: 0 milliseconds (values less than 1000 are ignored)

Options: Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see details)

Run Query  Reset
... and getting answers

<table>
<thead>
<tr>
<th>tool</th>
<th>name</th>
<th>toolType</th>
</tr>
</thead>
<tbody>
<tr>
<td>ark:/12148/br2d2cb</td>
<td>Magic mimeType Identifier</td>
<td>info:bnf/spar/agent#softwareAgent</td>
</tr>
<tr>
<td>ark:/12148/br2d2cb</td>
<td>Magic mimeType Identifier</td>
<td>info:bnf/spar/representation#identificationTool</td>
</tr>
<tr>
<td>ark:/12148/br2d238m</td>
<td>Outil JHOVE</td>
<td>info:bnf/spar/agent#softwareAgent</td>
</tr>
<tr>
<td>ark:/12148/br2d238m</td>
<td>Outil JHOVE</td>
<td>info:bnf/spar/representation#characterizationTool</td>
</tr>
<tr>
<td>ark:/12148/br2d2598</td>
<td>Xerces2 Java Parser 2.9.1</td>
<td>info:bnf/spar/agent#softwareAgent</td>
</tr>
<tr>
<td>ark:/12148/br2d2598</td>
<td>Xerces2 Java Parser 2.9.1</td>
<td>info:bnf/spar/representation#validationTool</td>
</tr>
<tr>
<td>ark:/12148/br2d26vb</td>
<td>Outil MediaInfo</td>
<td>info:bnf/spar/agent#softwareAgent</td>
</tr>
<tr>
<td>ark:/12148/br2d26vb</td>
<td>Outil MediaInfo</td>
<td>info:bnf/spar/representation#characterizationTool</td>
</tr>
<tr>
<td>ark:/12148/br2d28kt</td>
<td>Outil Jhove2</td>
<td>info:bnf/spar/agent#softwareAgent</td>
</tr>
<tr>
<td>ark:/12148/br2d28kt</td>
<td>Outil Jhove2</td>
<td>info:bnf/spar/representation#characterizationTool</td>
</tr>
<tr>
<td>ark:/12148/br2d2186</td>
<td>Outil File</td>
<td>info:bnf/spar/agent#softwareAgent</td>
</tr>
<tr>
<td>ark:/12148/br2d2186</td>
<td>Outil File</td>
<td>info:bnf/spar/representation#identificationTool</td>
</tr>
</tbody>
</table>
Updating the data: Named graphs

- ark:/12148/b1234.namedGraph
- ark:/12148/b1234.namedGraph.version0
- ark:/12148/b1234.namedGraph.version0.release0
... What next?
Some hints

• Publish part of our data online
  – Publish the **ontologies**
    • On data.bnf.fr with a specific sub-namespace
  – Publish our **instances**
    • All the reference information…
    • Especially the format and software registry

• Update with the state-of-the-art
  – From info:URIs to HTTP URIs
  – Stitch our ontologies with recent ones:
    • PREMIS, UDFR ontologies
  – And link to existing RDF data sets, e.g. PRONOM & UDFR
An ever-growing part of our bibliographic data is available as RDF at data.bnf.fr…

An ever-growing part of our digital assets have their preservation metadata expressed as RDF…
- Bridge bibliographic and preservation metadata?
- Thanks to ARK identifiers, this is feasible!

From a digital preservation tool to a collection management utility
Thank you for your attention

mailto:
sebastien.peyrard
@bnf.fr