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


SILVIE

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

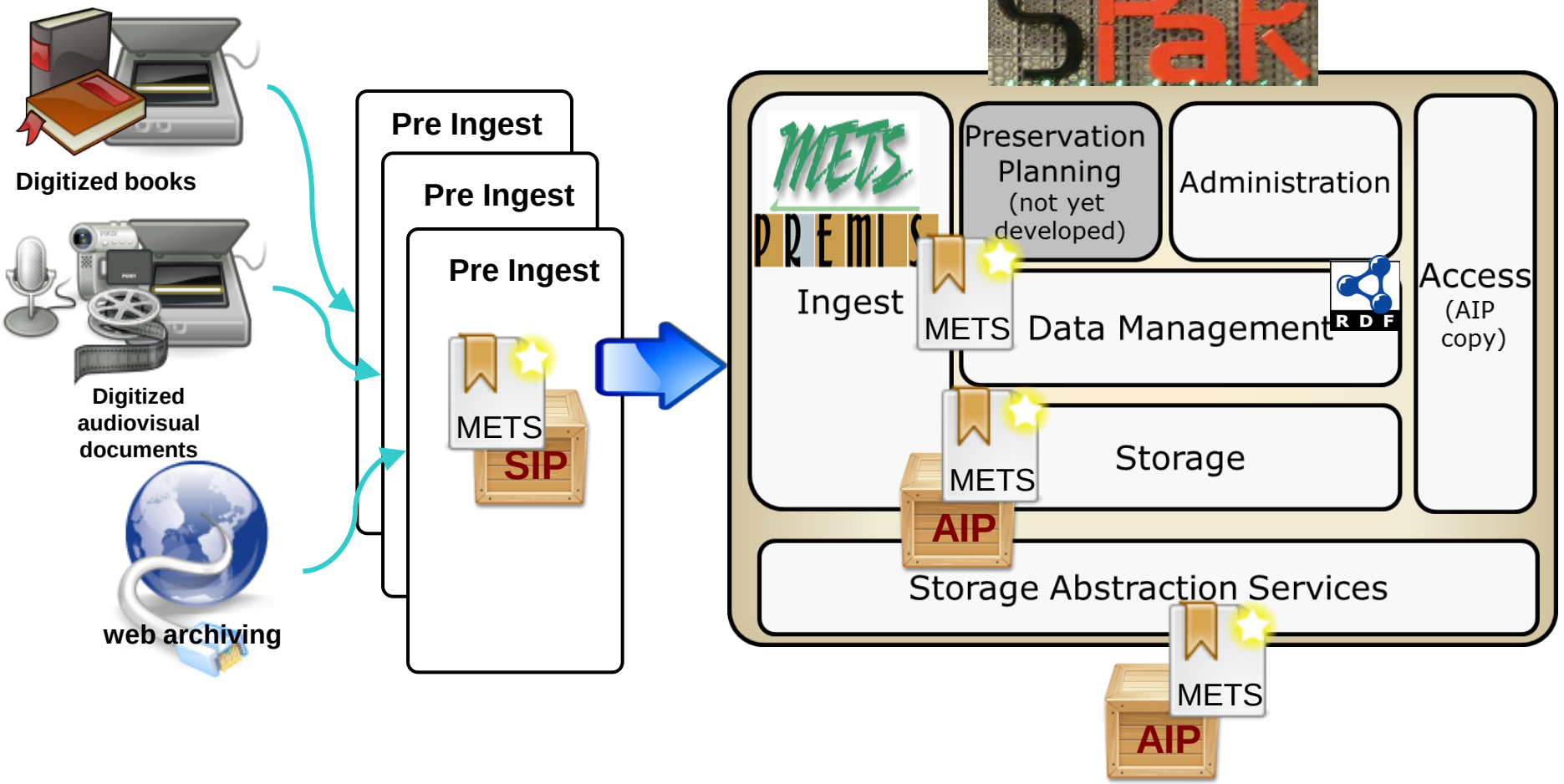
• a digital asset

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

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

• an authority record
• a bibliographic record

Preserving our assets the SPAR repository



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:

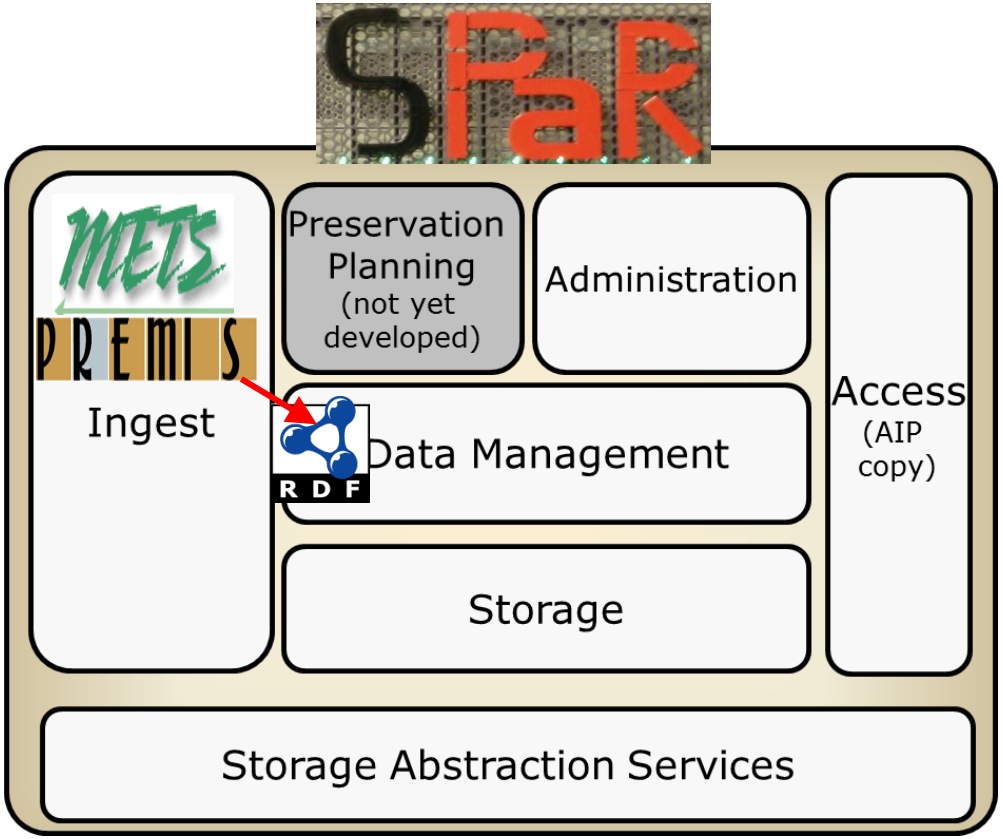
Needs unambiguous & persistent identifiers

An example :

```
<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/br2d2xn</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*

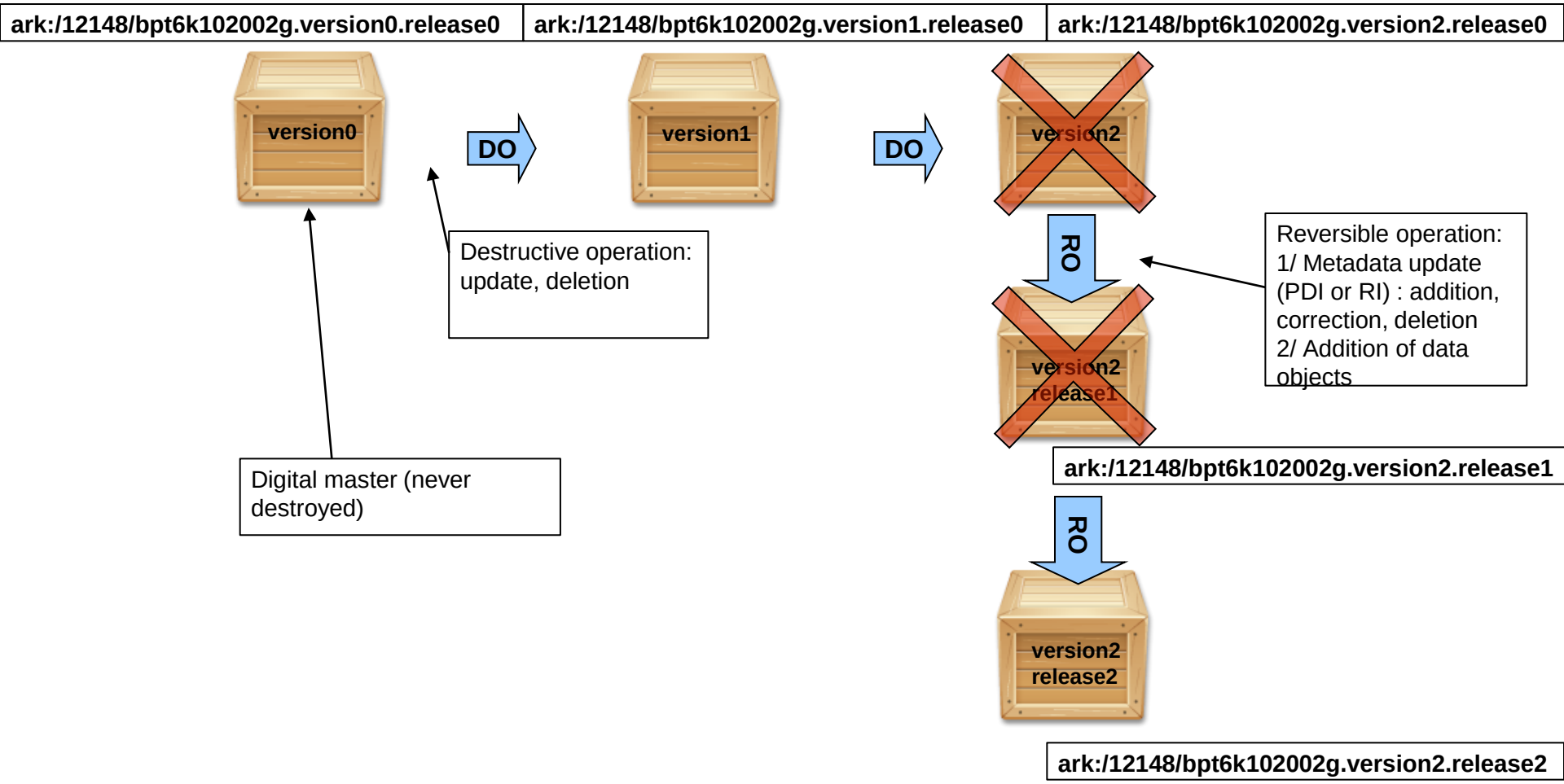
→ Need for a URI policy in SPAR

→ 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



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)

URIs for… abstract notions

- 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

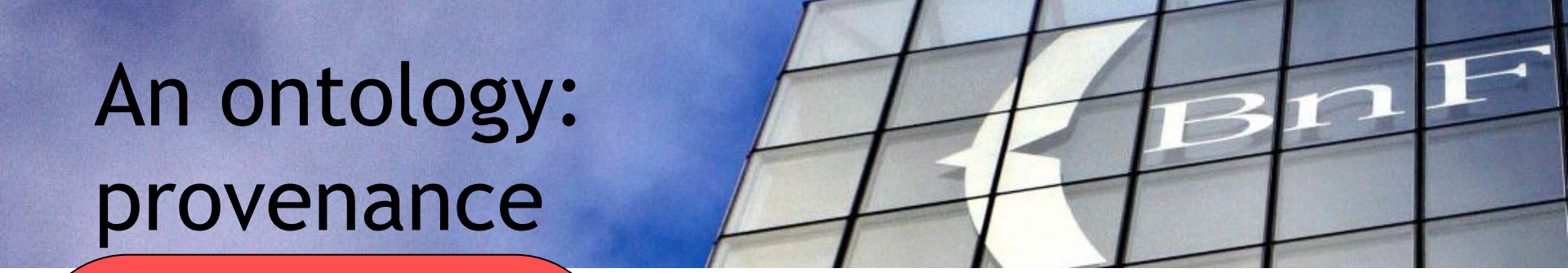
Designing the ontologies

- 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

DOAP

ORE

An ontology: provenance



http://purl.org/dc/dcmitype/Event

info:bnf/spar/provenance#**event**

- digitization
- idGeneration
- ocerization
- ingestCompletion
- fileProcessing
- packageUpdate
- etc.

http://purl.org/dc.elements/1.1/**date** → literal

info:bnf/spar/provenance#**hasAgent** → info:bnf/spar/agent#**agent**

hasPerformer

hasIssuer

info:bnf/spar/provenance#**hasProduct** → info:bnf/spar/provenance#**outcomeInformation**

info:bnf/spar/provenance#eventOutcome → literal

From METS-PREMIS to RDF

METS

```
<mets>
 <dmdSec ID="DMD.1">
  <mdWrap MIMETYPE="text/xml" MDTYPE="DC">
   <xmlData>
    <spar_dc:spar_dc>
     <dc:title>Impressions de Sicile</dc:title>
     <dc:creator>Volkonskaïa, Mariă</dc:creator>
    </spar_dc:spar_dc>
   </xmlData>
  </mdWrap>
 </dmdSec>
 <amdSec>
  <digiprovMD ID="AMD.25">
   <mdWrap MDTYPE="PREMIS:EVENT" MIMETYPE="text/xml">
    <xmlData>
     <premis:event>
      <premis:eventIdentifier>
       <premis:eventIdentifierType>UUID</premis:eventIdentifierType>
       <premis:eventIdentifierValue>
          f4ca87f0-8453-11df-8668-00144f68e1cc
       </premis:eventIdentifierValue>
      </premis:eventIdentifier>
      <premis:eventType>ingestCompletion</premis:eventType>
      <premis:eventDateTime>2010-06-30</premis:eventDateTime>
      <premis:eventDetail>ingestCompletion is valid</premis:eventDetail>
      <premis:linkingAgentIdentifier>
       <premis:linkingAgentIdentifierType>
          BnFApplication
       </premis:linkingAgentIdentifierType>
       <premis:linkingAgentIdentifierValue>
          ark:/12148/br2d27h/act09
       </premis:linkingAgentIdentifierValue>
       <premis:linkingAgentRole>performer</premis:linkingAgentRole>
      </premis:linkingAgentIdentifier>
     </premis:event>
    </xmlData>
   </mdWrap>
  </digiprovMD>
 </amdSec>
 </amdSec>
 <structMap TYPE="physical">
 <div TYPE="group" ID="DIV.2" DMDID="DMD.1" ADMID="AMD.25">
  <div TYPE="object" ORDERLABEL="NP" ORDER="1" ID="DIV.3">
   <fptr FILEID="master.1"/>
  </div>
 </div>
 </structMap>
</mets>
```

```
<ark:/12148/bpt6k206840w.version0.release0>
        rdf:type sparstructure:group> ;
        dc:title "Impressions de Sicile"  ;
        dc:creator "Volkonskaïa, Mariă" ;
        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" ;
        sparprovenance:hasPerformer
<ark:/12148/br2d27h/act09>.
```

SPARQL

query

```
SEL
WHE
?pac

?eve
      dc:date ?date ;
      sparprovenance:hasPerformer ?performer.    }
```

… and the answer is…

pac	.	digitization	performer
ark:/12148/bpt6k206840w.version0.release0	Volkonskaïa, Mariă	28/07/2010	ark:/12148/br2d27h/act09

Querying the data…

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
SELECT DISTINCT ?tool ?name ?toolType WHERE {
?file oai-ore:isAggregatedBy ?fileGroup.
?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

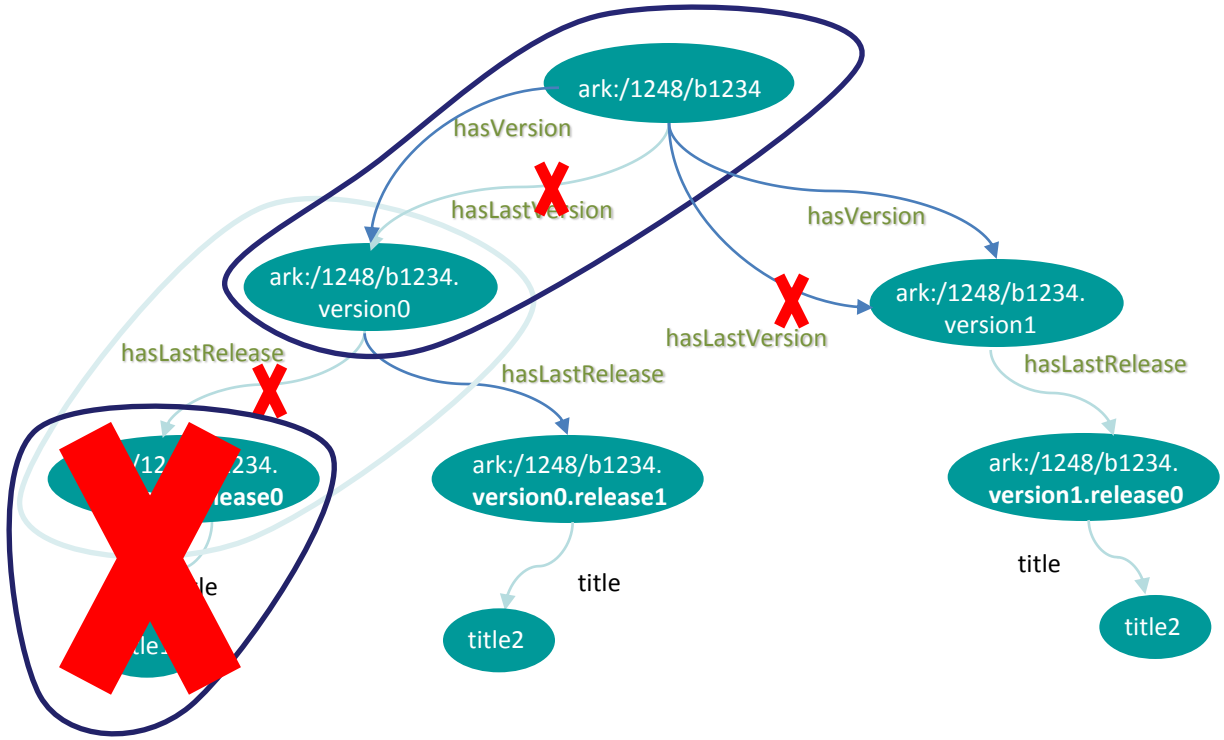
(The result can only be sent back to browser, not saved on the server, see details)

Run Query | Reset

… and getting answers

tool	name	toolType
ark:/12148/br2d2cb	Magic mimeType Identifier	info:bnf/spar/agent#softwareAgent
ark:/12148/br2d2cb	Magic mimeType Identifier	info:bnf/spar/representation#identificationTool
ark:/12148/br2d238m	Outil JHOVE	info:bnf/spar/agent#softwareAgent
ark:/12148/br2d238m	Outil JHOVE	info:bnf/spar/representation#characterizationTool
ark:/12148/br2d2598	Xerces2 Java Parser 2.9.1	info:bnf/spar/agent#softwareAgent
ark:/12148/br2d2598	Xerces2 Java Parser 2.9.1	info:bnf/spar/representation#validationTool
ark:/12148/br2d26vb	Outil MediaInfo	info:bnf/spar/agent#softwareAgent
ark:/12148/br2d26vb	Outil MediaInfo	info:bnf/spar/representation#characterizationTool
ark:/12148/br2d28kt	Outil Jhove2	info:bnf/spar/agent#softwareAgent
ark:/12148/br2d28kt	Outil Jhove2	info:bnf/spar/representation#characterizationTool
ark:/12148/br2d2186	Outil File	info:bnf/spar/agent#softwareAgent
ark:/12148/br2d2186	Outil File	info:bnf/spar/representation#identificationTool

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

One dream before the end

- 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