

# BnF and OCLC/OPF pilot

Why BnF is  
interested

Sébastien Peyrard

iPRES 2012



# Repository at BnF: up and running



Digitized books



Digitized audio and video



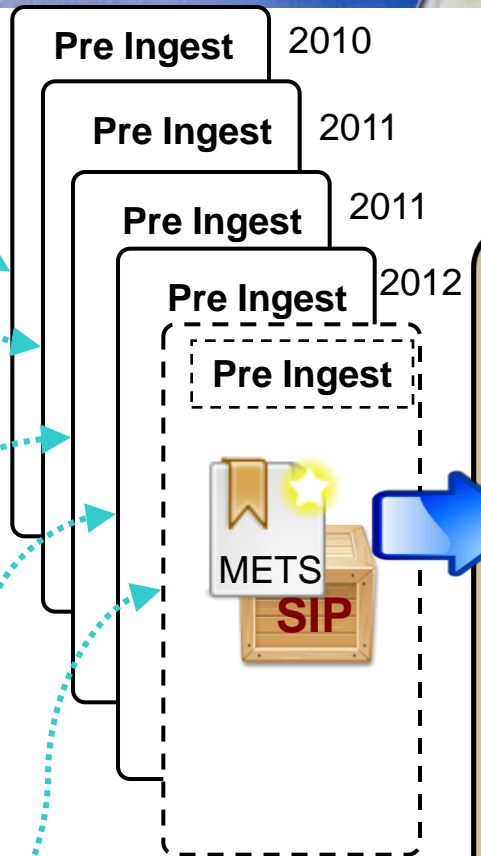
third party storage



web archives



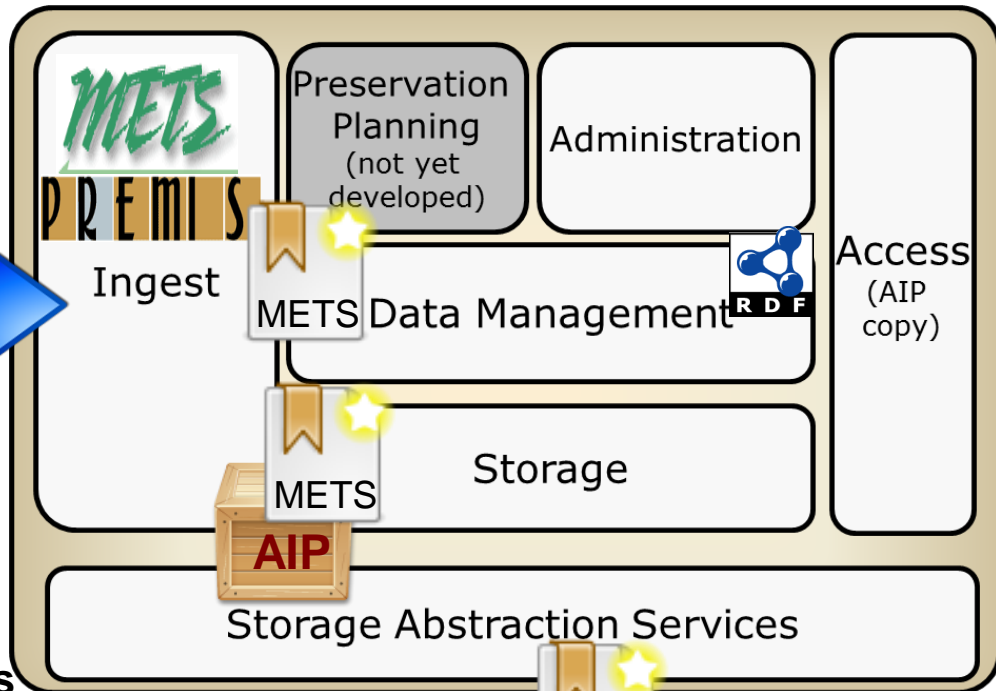
Archive records



450,000 ingested AIPs

97 million files

325 TB



# Preservation metadata at BnF



- 1 dedicated FTE about repository metadata with backup
- Traditional PREMIS in METS
- Mapped in an RDF triple store where it can be queried
- « Reference » information packages preserved, with structured, factorized metadata about
  - policies
  - formats
  - preservation tools
  - system processes

# Why pilot?

## 1. Obvious reasons

- Evaluate the quality of our metadata, places where it can be improved
  - Especially for a « data-first » system
- External evaluation is useful for changing
- Advocate for preservation metadata (secure existing activities)

## 2. Where the pilot meets our agenda



### BnF projects

- Integrate DP to the daily activities of the library: digital meets physical preservation
- Build new channels, where ensuring the mission continuity is crucial
  - Substitution legal deposit
  - Acquisitions
- Improve the data management interfaces
- Iterate on the risk analysis

### Pilot possible inputs

- Consider metadata from a risk driven approach
- Quantified need for pre-ingest provenance and context metadata
- Digital curator homepage risk table and priority requests
- Start with metadata

# Risk analysis: from this...



METS

PREMIS

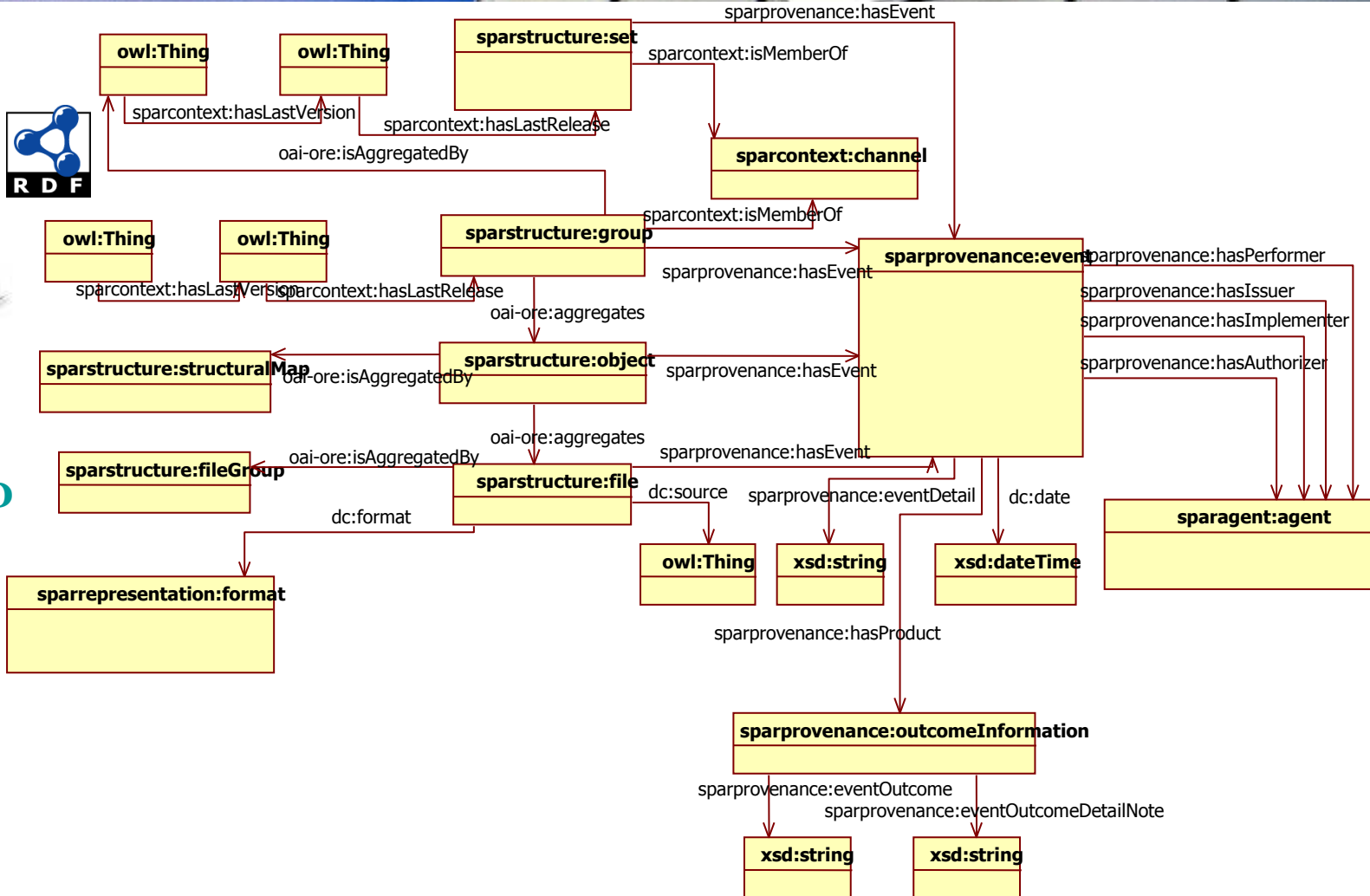


textMD

MIX

MPEG-7

containerMD



to something  
like this



Information1 (corresponding metadata field)→  
risk1 mitigated

Information2 (corresponding metadata field)→  
risk2 mitigated

→ « Functional requirements for preservation  
metadata »

- Explains why preservation metadata is important  
on a **concrete and quantifiable basis**
- Threat-driven culture talks to traditional curators  
and managers

# Curators don't want (just) this



## Virtuoso SPARQL Query Editor

[About](#) | [Namespace Prefixes](#) | [Inference rules](#)

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



# But something like that (too)!



<imagine user-friendly interface>

Risk analysis will help  
define what the most  
important questions are  
And organize the interface

<answers to request1>

<answers to request2>

<answers to request3>

# Questions?

sebastien DOT peyrard

AT bnf.fr

