From Several Datasets to One Graph

designing a prototype for
a graph visualization of archival metadata
from different institutions

Anila Angjeli - Bibliothèque nationale de France
Florence Clavaud - Archives nationales, France
Stéphanie Roussel - Service interministériel des Archives de France

Description Section – SAA Annual Conference, August 4, 2016, Atlanta
The Project in a few words
The participants (Who?)

- Three institutional partners:
  - Archives nationales, France (project lead)
  - Bibliothèque nationale de France
  - Service interministériel des Archives de France (SIAF)

- Emmanuel Château-Dutier – researcher
  from the University of Montreal
  A historian specializing in administration of public architecture projects
  in France and a digital humanist
The partners are joining their efforts to demonstrate the interest for interconnecting the archival resources under their care by using the semantic web technologies in the following way:

- **represent** in a rigorous way **in RDF** archival metadata (authority records, documents’ description) from several institutions;

- **interconnect** and enrich the resulting datasets, especially by expressing relationships that exist among the described objects;

- **visualize** the datasets in a useful way, graphically by enabling the user to **navigate** in the graph and access the source metadata

**A proof of concept and of utility**

Supported by the French Ministry of Culture and Communication as promoting the implementation of Semantic Web technologies in cultural institutions
How/Who?

- Project management:
  - project team – a limited number of persons, though others are involved
- Metadata preparation and RDF conversion - By the project team
- Prototype – By a contractor from the public or private sector (tender)
- Project co-funded by the three participating institutions
  - Global budget: €45,000 (for outside contractor)
  - Project team time contributed by the participating institutions
- The libraries realized over the project will be under open license

Timeline

- A three-year project 2015-2017
- Publication of the prototype – mid 2017
What’s at stake?
Institutional Context
Common objectives - Different stakes

National Archives

- 1998-2013
  - Significant effort to normalize and computerize archival descriptions and related creators: pilot implementations; IS development and production (encompassing all tasks of "archival chain"); "Virtual Finding Aids Room" – online

- Today
  - >22,000 FA (XML/EAD), 14,000 authority records on creators (XML/EAC-CPF); new FA each and every day; systematic description of creators

A complex network of relationships,
involving many types of objects, plus a temporal dimension

Problems:

1. The look-up and navigation tools for these data do not reveal the true dimensions of the network of relations

2. The AN are part of a large administrative and research data ecosystem in France
   But the metadata are not technically connected to those of other cultural heritage institutions

- First steps - Use of ISNI identifiers; alignment tests with BnF authority records…
Common objectives - Different stakes

Bibliothèque nationale de France

- A variety of repositories
  - General Catalogue
  - Archives & Manuscripts
  - ...
  - LOD exposure with Semantic web technologies

http://data.bnf.fr interconnecting resources from different repositories by using as connecting nodes the authority records

All metadata are under the French State Open Licence – EtaLab (similar to CC-By)

- On the side of the Administrative archives of the institution
  - Authority records in EAC-CPF – **outside the Archives & Manuscripts catalogue; no publication and visualisation tool**
  - Desire to **interconnect with other archival repositories**, starting with the National Archives (using ISNI identifier, among other means)
Common objectives - Different stakes

Service interministériel des Archives de France (SIAF)

supervises the network of public archives in France

- Manages a national, reference repository of patterns for the description of archives’ creators (in collaboration with the Association des Archivistes Français) – in EAC-CPF
  http://www.archivistes.org/Notices-d-autorite-producteurs-1781

  A "meta-repository" for interconnecting all institutions performing the same or similar function all over the territory (e.g.: Law Court of Justice)

- Desire to represent these authority descriptions in RDF and visualise the graph of record creators

- Desire to demonstrate the feasibility of interconnecting these descriptions with those of the Archives nationales (describing the archive’s creators of French central administration archives)
The perspective of a historian

• Evaluate the relevance and the efficiency of archival, documentary models
  • for the description of functions and corporate bodies

• Realize diachronic visualizations
  • to display and explore historical administrative, organizational charts
In summary

- Different situations
- A strategic project for the three institutional partners
- A real ambition
- A small but highly motivated team
- Strong interest in the methods and learnings of the project
- Results that might interest other entities
  (a lot of interest already expressed)
Qualitative objectives

• The prototype will be developed outside the Information System of each of the partners

• Selection of the dataset – relatively small volume
  • Reason: the purpose is not to process all the available data from each of the institutions., as this would need to take into account a great heterogeneity and envisage very robust tools

The project aims principally:

• **Quality of results**
  • Expressiveness of data
  • Relevance and accuracy in the expression of Relationships
  • Possibility of reasoning (inference)
  • Clarity
  • Ergonomic and reactive user interface

• **Rigorous methodological approach**
Technological context
the right moment!

- Reference ontology
- Data visualization technologies
An OWL ontology for archives

- Currently no standard ontology for archival description exists
  - Only four ICA standards

- In 2012 the ICA Experts Group on Archival Description (EGAD) began development of
  - a **conceptual model** (Records in Contexts: RiC-CM)
  - a **formal ontology** (RiC-O)

- First draft of RiC-CM to be presented at the ICA Conference, in September 2016, in Seoul

- First draft of RiC-O will be released in late 2016
A very simple example of a graph that RiC-O could help to build
Use RiC-O for the prototype

• Florence Clavaud at the ANF and leader of this project is coordinating the development of RiC-O within EGAD

• Metadata selected for the project describe objects that represent a subset of the entities defined by RiC-O

  • Instances of the classes
    • Corporate Body
    • Person
    • Record Set
    • Activity
    • …

• **Relation between our project and RiC-O**
  • The project (if all goes well) might serve as a use case for RiC-O
  • Provide feedback to RiC-O
Data visualization on the web

• Technologies and libraries exist
  • Base languages: JavaScript, AJAX, JSON, HTML5
  • data visualization (as D3)
  • generation of timelines (e.g.: Timeline of SIMILE project)
  • Other developments are underway in digital humanities based on existing libraries

• But nothing is directly reusable

• Our problem is complex because the system to be represented is complex
  • Different types of “objects” (CPF, Functions/Activities, Records and Records Sets other types of documents, Dates, etc.)
  • Different types of relations
  • Density of relations
  • Different institutional perspectives to be combined
  • …
Selection and preparation of source metadata
Definition of metadata sets

• The contributed dataset from each of the partners should be:
  • relevant
  • controllable (amount, characteristics…)
  • possible to interconnect with the data provided by the others

• Selection of two functional domains
  • Management of historical monuments and “civil buildings”
  • Public reading

  These include:
  • Entities that are part of the governmental (central) administration
  • National operators
  • Territorial services

  Temporal coverage:
  • Contemporary period

  Described objects:
  • Actors of the domain, the records created, the functions, etc.
Method

• Point of departure: authority records on archives’ creators (including both Corporate Bodies and Persons)
  • A consistent body of records, carefully curated

• Each of the partners preserves its own perspective (no specific common rules in building the records)

However, agreement on:
• Encoding format - XML / EAC-CPF
• Form of the heading – respect of the French cataloguing rules
• Expression of legal statuses, functions and their indexing
• Establishment of relationships and their types (RiC-O)
Where are we today? (preparation of the metadata set)

- National Archives
  - Domain of Historical monuments - descriptions of around 40 entities
  - Domain of public libraries – descriptions of around 60 entities (services of the central administration, related operators, some persons) from the 70’s

- BnF
  - Domain of public libraries – descriptions of around 55 entities (mainly describing the services of the BnF from the Ancient Regime period)
  - Domain of Historical monuments – descriptions of around 15 entities related to the architectural projects of the BnF buildings

- SIAF
  - Descriptions of around 100 entities (territorial services on both functional domains)

- The Researcher (historian)
  - Descriptions of around 15 entities (domain of Historical monuments and their architectural history)
First findings
Overlaps and Complementarity

• Significant overlap between the entities described by the BnF and by the Archives nationales

Include
• Personalities (secretaries, presidents of the BnF, directors of the Archives Nationales, architects, etc.)
• Administrative ecosystem of the BnF (Ministry of Culture, other partners) partially overlaps with the graph of archives’ creators described by the Archives Nationales.

• Complementarity
• The network of the administrative entities of the BnF is an extension of the central administration entities, as viewed by the Archives Nationales
• The network of the State administration entities for the cultural domain, as viewed an described by the Archives Nationales, specifies the ecosystem of the BnF partners
• The ecosystem of decentralized territorial corporate bodies described by SIAF is compliant with that of the Archives Nationales

• Historical dimension
• Strong, both in the BnF and in the Archives Nationales datasets (of course they reflect the historical dimension of archival holdings themselves)
Provisional conclusion

Interconnecting the datasets is a relevant objective

Confirmed by the analysis of the metadata sets and despite the different perspective of each of the partners

A requirement: keep track of the provenance of the source information
Further: Focus on relations
Modeling relations

- Expressing relations with high precision is essential
  - to meaningfully navigate the graph
  - to understand how documents created by the represented entities are complementary

- Relations defined by ISAAR(CPF), EAC-CPF and ISDF are not sufficient

- Currently the project
  - relies on the Archives Nationales’ guidelines for the creation of authority records (under development)
  - follows and is inspired by the work being done for the development of RiC-CM and RiC-O
Relationships between corporate bodies and persons

• Useful relationships for the project:
  • is Member Of
  • is Affiliated To
  • is Employee Of
  • is Director Of

• works for (?)
• influences / is Influenced By (?)
Relationships between corporate bodies

- Hierarchical relationships:
  - Whole/Part relationship
    - to meaningfully navigate in the graph
    - to understand how documents created by the represented entities are complementary

- Relationships defined by ISAAR(CPF), EAC-CPF and ISDF are not sufficient

- Currently the project
  - relies on the Archives Nationales’ guidelines for the creation of authority records (under development)
  - follows and is inspire by the work being done for the development of RiC-CM and RiC-O
Scope statement and specifications
Some core requirements

- Web application in compliance with standards and best practices
- Input data analysis
- Open source
- Architecture based on an RDF triplestore
- Functionalities
  - Alignments
  - SPARQL queries
  - Construction of new triples
  - Analytical cartographies
  - Graph navigation
  - …
Final (provisional) thoughts

Beyond simply building a prototype, the project should be instructive in terms of methodology

• How to build an inter-institutional project where each of the institutions is equally represented and participates actively?

• How to work together on the quality of our data, both:
  • In a coherent and unified way for the needs of the project,
  • And make each of the institutions benefit more broadly from this qualitative work?
Thank you!

(more next year!)

Anila Angjeli - Bibliothèque nationale de France
Florence Clavaud - Archives nationales, France
Stéphanie Roussel - Service interministériel des Archives de France

Description Section – SAA Annual Conference, August 4, 2016, Atlanta