ARCHIVES NATIONALES

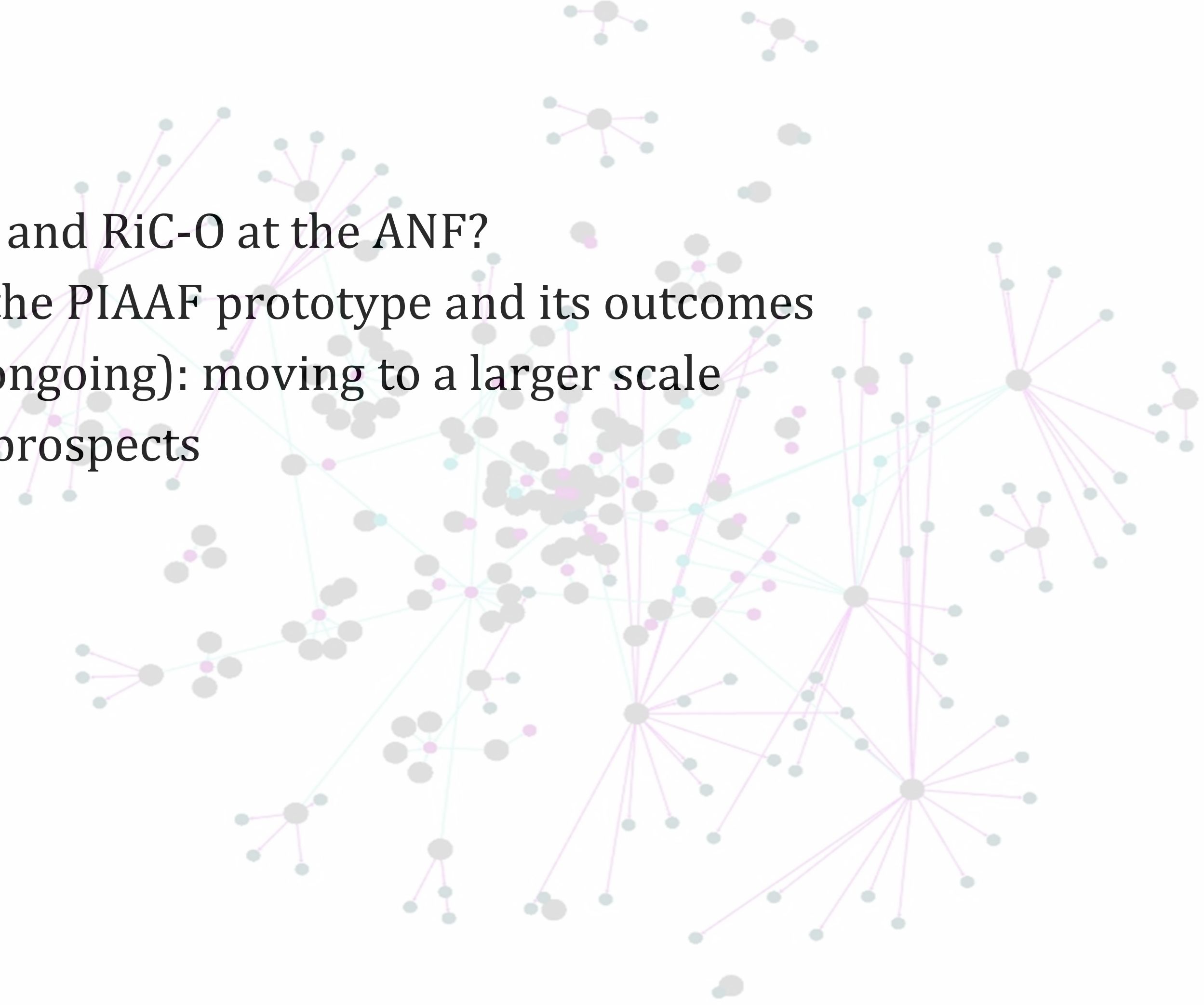
Using Records in Contexts (RiC) on a large scale: a feedback on generating, querying and displaying big, « real world », archival metadata sets conforming to RiC Ontology

Florence Clavaud, Pauline Charbonnier
(Archives nationales de France, Lab des Archives)

Friday, October 1st, 2021

ICA Virtual Conference 2021 –Empowering Knowledge Societies

Outline



- Why use RiC and RiC-O at the ANF?
- A first step: the PIAAF prototype and its outcomes
- Next steps (ongoing): moving to a larger scale
- Conclusion: prospects

A few words about
the Archives nationales de France (ANF)

- A national public service created on 1 January 2007 by a decree issued by the Ministry of Culture (dated 24 December 2006). Its mission is to collect, arrange, describe, preserve, provide access to, and promote public **archives from the central administrations of State, the archives of Parisian notaries, and private records of national interest**.
- In fact, **an old institution, created as a result of the French Revolution; records kept are dated from the Middle Ages to nowadays.**
- **Key figures** (2020) (see also https://www.archives-nationales.culture.gouv.fr/en/web/guest/chiffres-cles):
  - more than 370 linear km of analogue records, 70 To of digital archives
  - about 8,8 million records digitized and accessible online
  - 8000 different readers, about 115 000 boxes accessed in 2019
- An institutional website (https://www.archives-nationales.culture.gouv.fr/), in which you can query and display the archival metadata that describe the holdings, using the Salle des inventaires virtuelle (SIV) (https://www.siv.archives-nationales.culture.gouv.fr)

Archival metadata at the ANF today

- **A huge amount of heterogeneous metadata**, created by generations of archivists and historians over centuries
  A significant effort for digitizing them from 1990, and for updating and completing this legacy.

- In a **series of silos**; the main one contains more than **29000 archival finding aids (XML/EAD files, thus structured documents) and about 15000 authority records (XML/EAC-CPF files)** on the archival creators, as well as about 20 vocabularies in a specific XML format.
  But it is not the only one… We have a lot of databases, and other repositories (among which a digital library, and the recently implemented digital archival system).

Archival metadata at the ANF today: already a graph

- **Already a lot of relations between the files in the first silo, but not really viewable and not searchable through the front-end web application** (the Salle des inventaires virtuelle) and, as concerns the relations between the finding aids and the authority records, of one category only (the provenance relation)
  In the table, the relations are counted once, the inverse ones are not taken into account.

- The **vocabularies and other authority records of the main silo can be used for indexing the finding aids**, but are rather **poor**, **not standardized and quite rarely used till now.**

15 210 authority records in January 2021

Relation type	Number of relations
Hierarchical relation	9610
Temporal relation	4310
Associative relation	6036
Family relation	499
Total	20455

Statistics of relations between the authority records that describe the producers

Archival metadata at the ANF today: user perspective

- Several end-user interfaces

- **Very few intuitive access points (who, when, where, what...)**

- **A lot of redundancies, from one silo to another and also within the same silo** (particularly the main one, where the same group of records or same record may have been described several times in several finding aids)

- **The end user finds it difficult to understand the result lists and how the results are displayed (in the context of a finding aid, as subcomponents of it)**

- Very few bridges to other information systems

> In fact, a classic situation, at a huge scale

Needs, and a hypothesis

- Need to:
  - **Improve the consistency and granularity of the metadata, and facilitate their management**
  - **Break the internal silos**, link the data sets to each other through **shared authority data**
  - **Improve the access to metadata**, develop a new unique web interface
  - Move towards interoperability with external datasets
- The hypothesis: RiC can help to do this:
  - RiC-CM, as the global, conceptual, reference framework for designing the future graph of linked entities
  - RiC-O, as the technical model for building RDF knowledge graphs from the existing metadata and exposing them as Linked Open Data

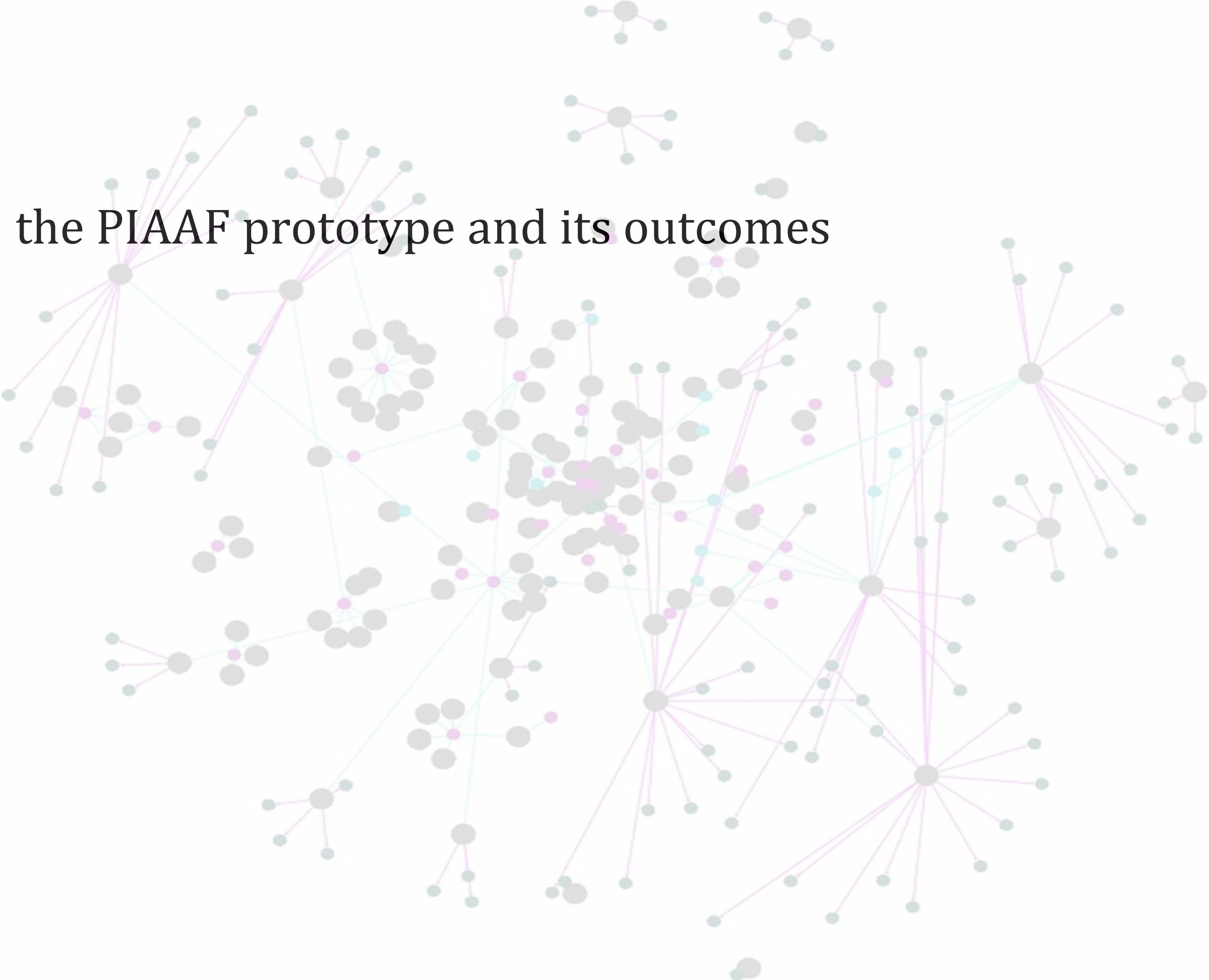
This assumption has become a certainty over the first achievements.

What RiC can help to do (a small, not exhaustive list): a global shift

- Focus on the objects that are to be described
- Make the relations between archival aggregations, or between records, explicit
  reveal the underlying graph
- Consider the finding aids as... records
- Consider the digital images of analogue records as physical instantiations (among others) of a unique intellectual content
- Consider when needed some curation events, and more generally speaking, the history of the records
- Consider other relations between records and agents than the provenance relation
- Consider places as geo-historical entities
- Etc.

Outline

- A first step: the PIAAF prototype and its outcomes



A first proof of concept: PIAAF

- PIAAF : « Pilote d'Interopérabilité pour les Autorités Archivistiques Françaises »
- A prototype to prove that:

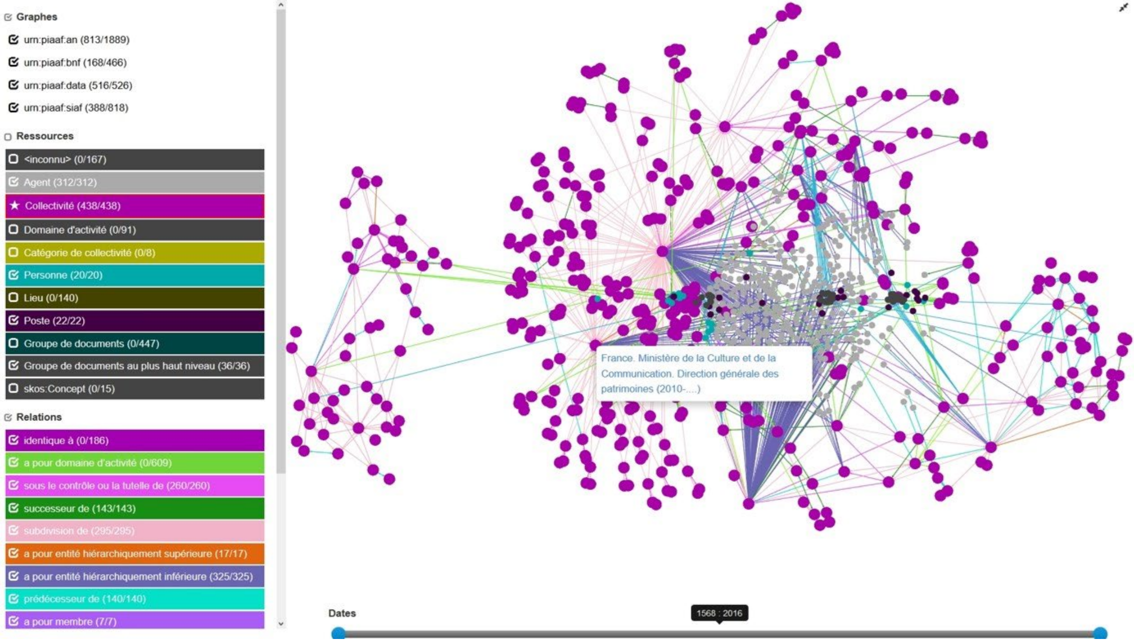
  - real-world existing (XML/EAD and XML/EAC-CPF) archival metadata sets can be converted to good quality RDF/RiC-O datasets, without losing any information;

  - RDF/RiC-O datasets coming from several institutions can be interconnected and linked to other datasets;

  - new web interfaces can be built for these sets, that enable to query, visualize and browse the graph thus generated
- A collaborative project lead by the ANF, with the French National Library and the Ministry of Culture, and a French private company, Logilab
- Qualitative project, on a small quantity of metadata, using an early version of RiC-O
- Project started in 2015; datasets selected, then accurately checked; converted to RDF by the ANF; prototype published online in February 2018 : https://piaaf.demo.logilab.fr

PIAAF interface: one can see the underlying graph

A screenshot from the Corporate bodies (Collectivités) tab (https://piaaf.demo.logilab.fr/ric/CorporateBody), once other entities selected as well as the relations between these entities



☑ **Graphes**

- ☑ urn:piaaf:an (813/1889)
- ☑ urn:piaaf:bnf (168/466)
- ☑ urn:piaaf:data (516/526)
- ☑ urn:piaaf:siaf (388/818)

□ **Ressources**

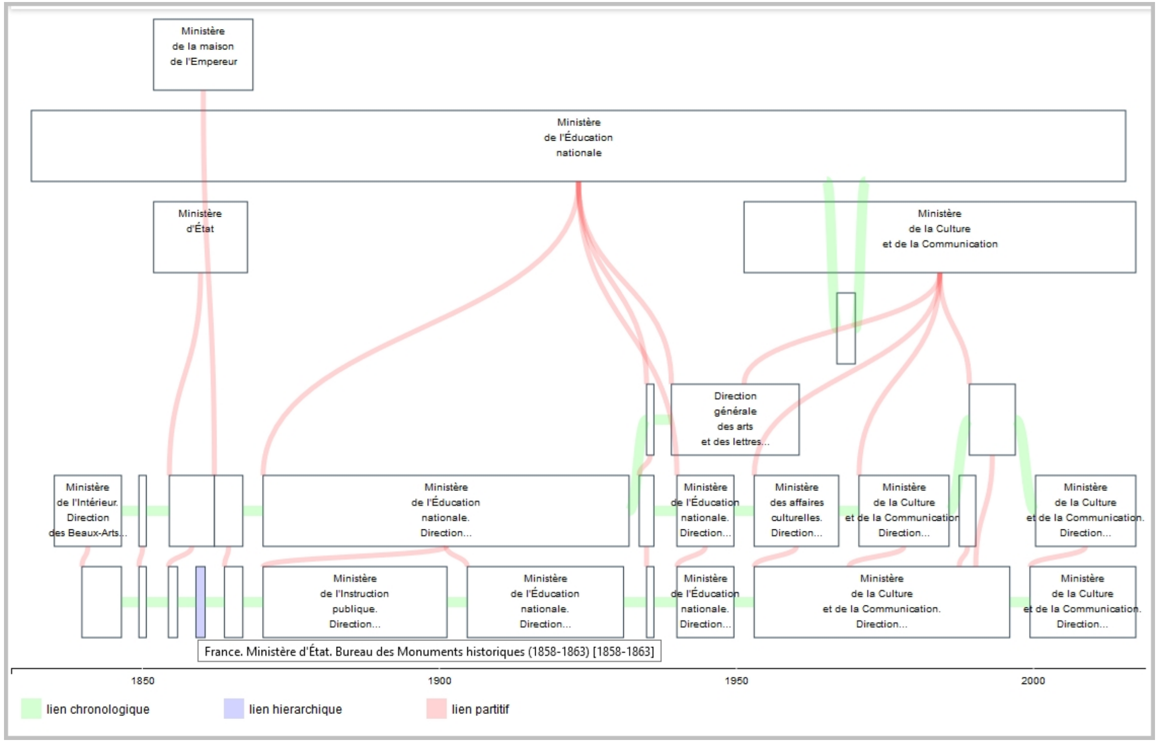
- □ <inconnu> (0/167)
- ☑ Agent (312/312)
- ★ Collectivité (438/438)
- □ Domaine d'activité (0/91)
- □ Catégorie de collectivité (0/8)
- ☑ Personne (20/20)
- □ Lieu (0/140)
- ☑ Poste (22/22)
- □ Groupe de documents (0/447)
- ☑ Groupe de documents au plus haut niveau (36/36)
- □ skos:Concept (0/15)

☑ **Relations**

- ☑ identique à (0/186)
- ☑ a pour domaine d'activité (0/609)
- ☑ sous le contrôle ou la tutelle de (260/260)
- ☑ successeur de (143/143)
- ☑ subdivision de (295/295)
- ☑ a pour entité hiérarchiquement supérieure (17/17)
- ☑ a pour entité hiérarchiquement inférieure (325/325)
- ☑ prédécesseur de (140/140)
- ☑ a pour membre (7/7)

France. Ministère de la Culture et de la Communication. Direction générale des patrimoines (2010-....)
Dates
1568 : 2016

PIAAF: an example of an automatically generated chart (« diagramme chronologico-hiérarchique »)


Ministère de la maison de l'Empereur
Ministère de l'Éducation nationale
Ministère d'État
Ministère de la Culture et de la Communication
Direction générale des arts et des lettres...
Ministère de l'Intérieur. Direction des Beaux-Arts...
Ministère de l'Éducation nationale. Direction...
Ministère de l'Éducation nationale. Direction...
Ministère des affaires culturelles. Direction...
Ministère de la Culture et de la Communication. Direction...
Ministère de la Culture et de la Communication. Direction...
Ministère de l'Instruction publique. Direction...
Ministère de l'Éducation nationale. Direction...
Ministère de l'Éducation nationale. Direction...
Ministère de la Culture et de la Communication. Direction...
Ministère de la Culture et de la Communication. Direction...
France. Ministère d'État. Bureau des Monuments historiques (1858-1863) [1858-1863]
1850
1900
1950
2000
lien chronologique
lien hierarchique
lien partitif

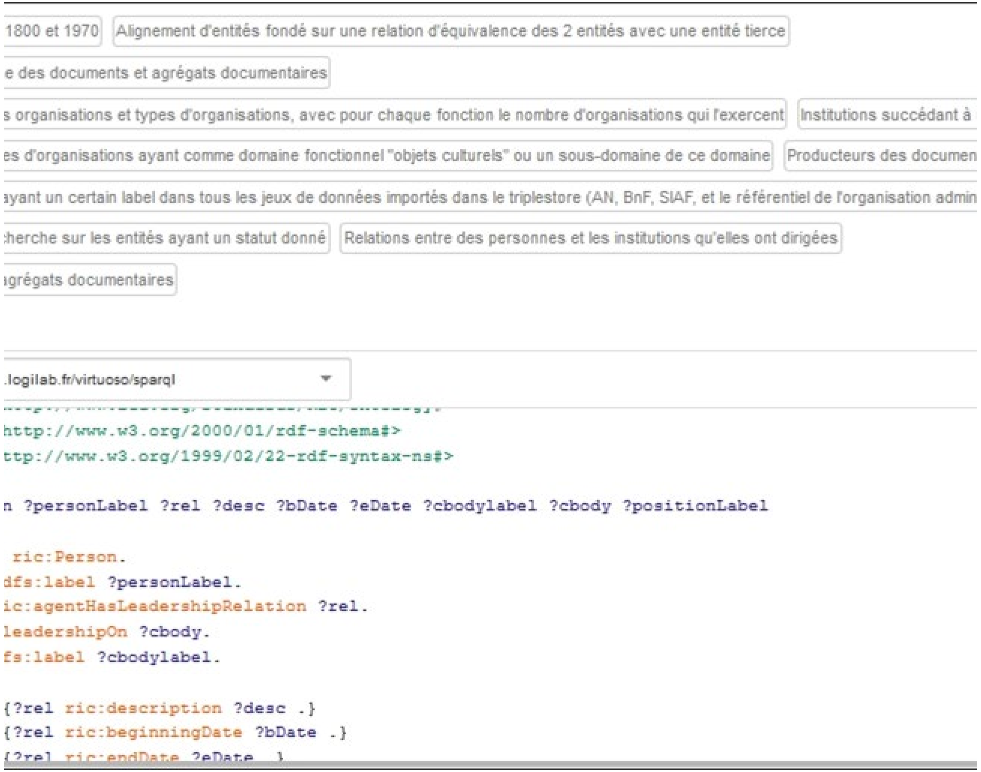
A partial copy of the web page on the corporate body whose name is « France. Ministère d'État. Bureau des Monuments historiques (1858-1863) » (https://piaaf.demo.logilab.fr/resource/FRAN_corporate-body_051123)

Using SPARQL for querying a RDF/RiC-O graph

Screenshot from the page https://piaaf.demo.logilab.fr/sparql. The SPARQL endpoint includes previously recorded queries.

The prototype also includes :

- A tutorial
- A tab showing data linking between French national archives and French national library data.
- The project documentation.



1800 et 1970 | Alignement d'entités fondé sur une relation d'équivalence des 2 entités avec une entité tierce

e des documents et agrégats documentaires

s organisations et types d'organisations, avec pour chaque fonction le nombre d'organisations qui l'exercent | Institutions succédant à

es d'organisations ayant comme domaine fonctionnel "objets culturels" ou un sous-domaine de ce domaine | Producteurs des documen

ayant un certain label dans tous les jeux de données importés dans le triplestore (AN, BnF, SIAF, et le référentiel de l'organisation admin

cherche sur les entités ayant un statut donné | Relations entre des personnes et les institutions qu'elles ont dirigées

agrégats documentaires

.logilab.fr/virtuoso/sparql

```
http://www.w3.org/2000/01/rdf-schema#>
ttp://www.w3.org/1999/02/22-rdf-syntax-ns#>

n ?personLabel ?rel ?desc ?bDate ?eDate ?cbodylabel ?cbody ?positionLabel

 ric:Person.
dfs:label ?personLabel.
ic:agentHasLeadershipRelation ?rel.
leadershipOn ?cbody.
fs:label ?cbodylabel.

{?rel ric:description ?desc .}
{?rel ric:beginningDate ?bDate .}
{?rel ric:endDate ?eDate .}
```

PIAAF: the project results

- Project successful! A lot of not trivial work, including converting the data to RDF, but its outcomes were high quality ones, and beyond them **the methodology proved to be of great help for the next steps**.
- A very encouraging experience of an implementation of RiC-O, which confirmed that:

  - RiC-O can be used for expressing accurately the complexity of the world of archives

  - **interesting outcomes can be generated from real-world existing metadata**. However, **several aspects must be carefully taken into account**:

    - the existing information systems and the data models and processes they use have to be assessed, and if necessary changed, in order **to manage unique identifiers for the objects that may emerge;**

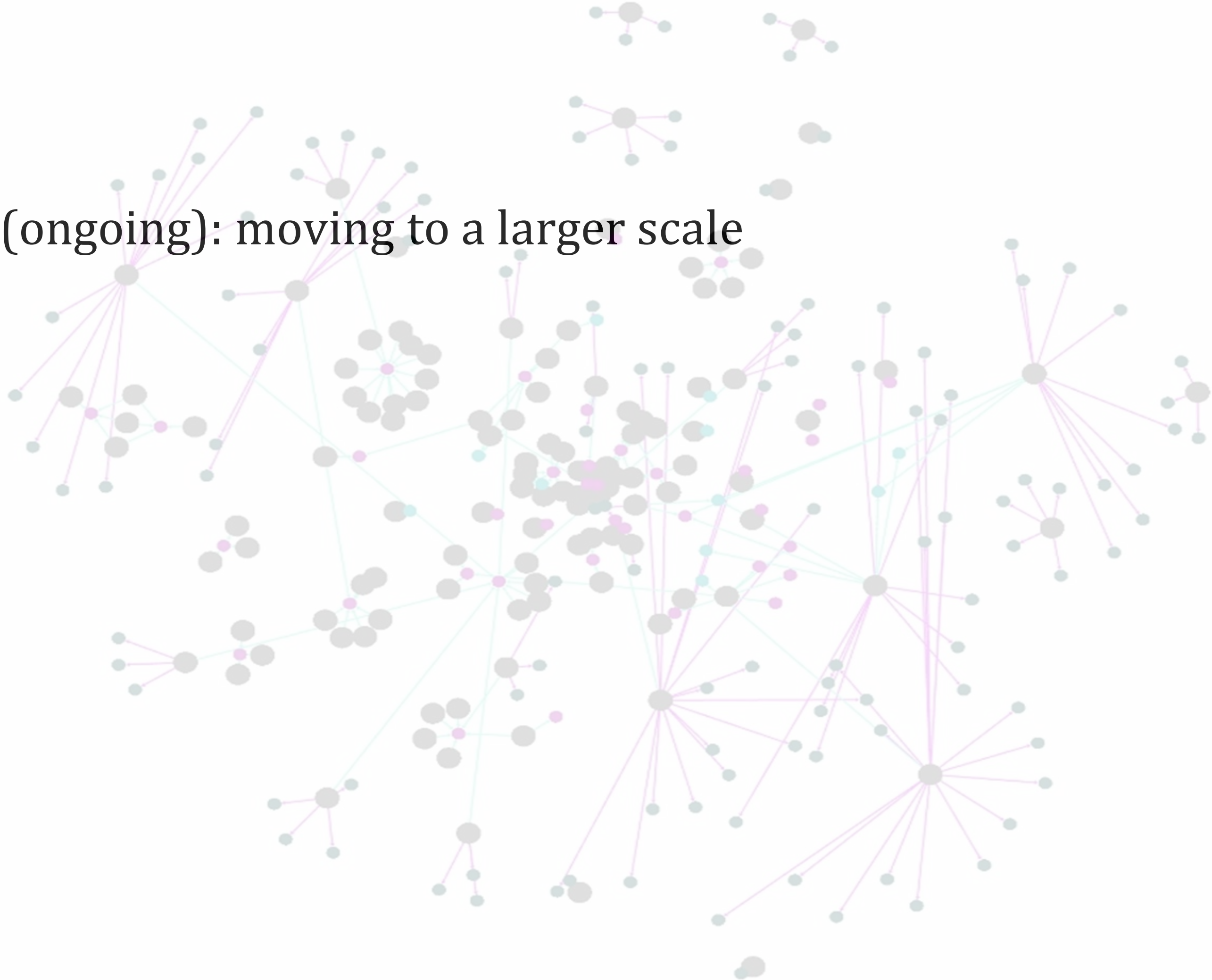
    - and more generally speaking, **the quality of the source metadata could be enhanced and more accurately controlled**

    - as concerns **the visualisation of the graph and the queries, very interesting new pathes had been opened, and could be further investigated**

- **An issue: PIAAF has not been updated since Feb. 2018**

Outline

- Next steps (ongoing): moving to a larger scale



Next (ongoing) steps: moving forward to a far larger scale

- Generating RDF sets from larger metadata sets

  - from the whole of the ANF finding aids and authority records (2019-2020): developing a software, RiC-O Converter

  - from the whole of the other authority data (vocabularies) (ongoing): defining a new, richer, SKOS and RiC-O based target data model, converting the source files to data conforming to this data model, and **enriching the resulting datasets**

  - from the other metadata silos (from 2021): database per database, and also from the digital archival system

RiC-O Converter software

- A software for **converting into RDF/RiC-O datasets all the EAD and EAC-CPF files held by the ANF**
- Developed from March 2019 to April 2020, by Sparna and the ANF; funded by the Ministry of Culture
- **Takes into account various scenarios, since the granularity of the files, and sometimes the content model of the elements, change from one file to another, or within the same file**
- **A strategy had to be defined and applied in order to deduplicate the relations between agents, and to generate unique persistent identifiers when the source files did not provide them**
- **Today the ANF can convert into RiC-O sets, using this software, the most important set within their metadata system, in less than 30 min on a personal computer.**
  The process generates about 155 million RDF triples for now.

RiC-O Converter: an open source, configurable and adaptable, software

- **Source code and documentation (in English), as well as unit tests, are available at https://github.com/ArchivesNationalesFR/rico-converter**

- The software can easily be configured, and is written mainly in XSLT 2 (; a of XSLT stylesheets are encapsulated in a Java script

- RiC-O Converter is **released under a free open source license (BSD-like)**.
  It can be reused and modified by any person or institution.

- **It comes along with EAD 2002 to RiC-O and EAC-CPF to RiC-O mappings, that can be used as a basis for developing any other tool of the kind.**

RiC-O Converter: more information

- The roadmap includes:
  - Developing a version 2.0 of RiC-O Converter that conforms to RiC-0 0.2
  - If we can find some funding, extending the scope of RiC-O Converter to the XML/SEDA files, used in French digital archives management systems to describe these digital archives.

- For more information on the tool and the outcomes of the project:
  - The video recording of the webinar dated 2020, June 20th (duration: about 2h; language: French): https://www.dailymotion.com/playlist/x6x1d0
  - The presentation dated December 2020 for the SemWeb Pro 2020 event (in French): https://peertube.semweb.pro/videos/watch/c195e540-4691-4b38-9619-0a59300d1adc;
  - And a recent article: Francart (Thomas), Clavaud (Florence), Charbonnier (Pauline). *RiC-O Converter: a Software to Convert EAC-CPF and EAD 2002 XML files to RDF Datasets Conforming to Records in Contexts Ontology*, in Proceedings of the Linked Archives international workshop, September 13th, https://drive.google.com/file/d/1mZoYjBCdjOqUZddgBRhWeEeDY-MHwjy8/view?usp=sharing

Converting to a new, richer, data model, and enriching, the vocabularies and authority data

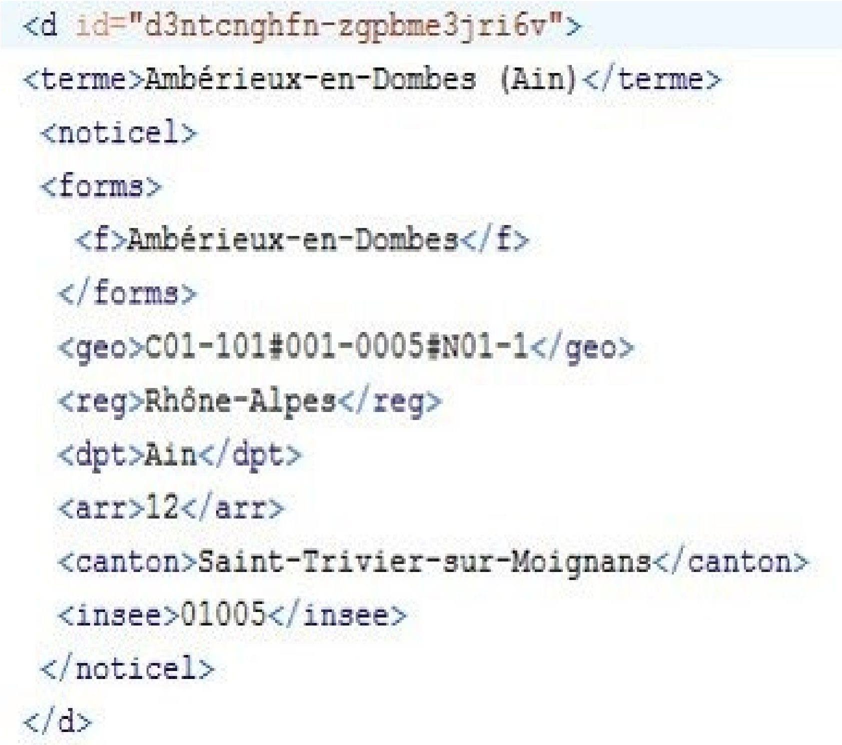
- For now, we have converted the ANF 20 vocabularies, whose source files had a very simple, most often flat, XML structure, to RDF/SKOS and RiC-O files.

- The source files, fortunately, assign a unique identifier to each concept or entity described.

- We are now enriching them, aligning them to other RDF authority data, or using other resources.

- We have also begun to build new authority data, the main one being a vocabulary on the corporate bodies' activity domains and processes. This project is a collaborative one.

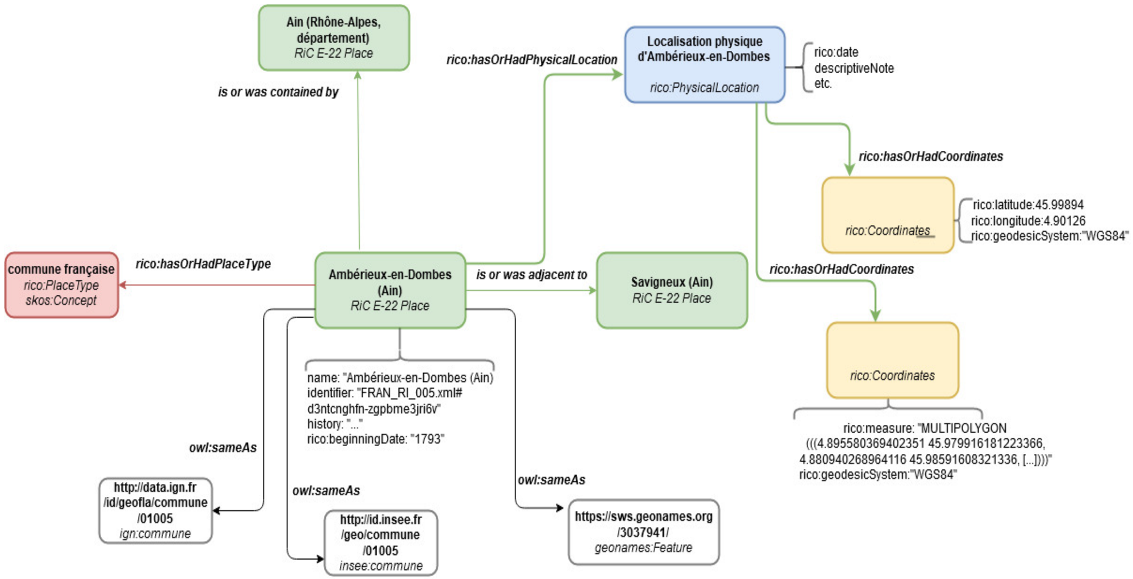
A specific example: working on the description of places



```
<d id="d3ntcnghfn-zgpbme3jri6v">
<terme>Ambérieux-en-Dombes (Ain)</terme>
 <noticel>
 <forms>
   <f>Ambérieux-en-Dombes</f>
  </forms>
  <geo>C01-101#001-0005#N01-1</geo>
  <reg>Rhône-Alpes</reg>
  <dpt>Ain</dpt>
  <arr>12</arr>
  <canton>Saint-Trivier-sur-Moignans</canton>
  <insee>01005</insee>
 </noticel>
</d>
```

A French local administrative area (the ‘commune d’Ambérieux-en-Dombes’) in the source file that describes the places in the ANF information system for now.

A specific example: places in RiC-O are considered as geo-historical entities



Ain (Rhône-Alpes, département)
RiC E-22 Place

is or was contained by

rico:hasOrHadPhysicalLocation

Localisation physique d'Ambérieux-en-Dombes
rico:PhysicalLocation

rico:date
descriptiveNote
etc.

rico:hasOrHadCoordinates

rico:Coordinates

rico:latitude:45.99894
rico:longitude:4.90126
rico:geodesicSystem:"WGS84"

commune française
rico:PlaceType
skos:Concept

rico:hasOrHadPlaceType

Ambérieux-en-Dombes (Ain)
RiC E-22 Place

is or was adjacent to

Savigneux (Ain)
RiC E-22 Place

rico:hasOrHadCoordinates

rico:Coordinates

name: "Ambérieux-en-Dombes (Ain)
identifier: "FRAN_RI_005.xml#
d3ntcnghfn-zgpbme3jri6v"
history: "..."
rico:beginningDate: "1793"

rico:measure: "MULTIPOLYGON
(((4.895580369402351 45.979916181223366,
4.880940268964116 45.98591608321336, [...])))"
rico:geodesicSystem:"WGS84"

owl:sameAs

http://data.ign.fr
/id/geofla/commune
/01005
ign:commune

owl:sameAs

http://id.insee.fr
/geo/commune
/01005
insee:commune

owl:sameAs

https://sws.geonames.org
/3037941/
geonames:Feature

Description partielle, conforme à RiC-O,
d'un lieu (la commune d'Ambérieux-en-Dombes) dans la version RDF du référentiel des lieux des Archives nationales.

Les relations sont présentées uniquement dans un sens (les relations inverses existent mais ne figurent pas dans le diagramme).
Le préfixe rico: indique que le composant (classe ou propriété) n'existe que dans RiC-O.
Lorsqu'il n'y a pas de préfixe, le composant existe dans RiC-CM et dans RiC-O.

- This is the graph resulting from processing and enriching (through alignments) the legacy data
- We also plan to add links to the name(s) of the place, to places that preceded it, or that are located within it.
- This place can further be linked to agents of which it is or was the location or jurisdiction, or record resources that have subject the place, etc.

Working on the authority data: first results and outcomes

- **Some of the RDF datasets built are already used by other projects**
  Example: the RDF sets describing places, and RDF sets created from a few finding aids, are being used by the ALEGORIA research project (web platform to be released in 2021).
- **A lot of work still to be done, and need to collaborate with other projects and institutions**
- Though all this is done out of the current information system, **we now have a standardized version of these authority data, and this work enabled us to begin to define a new model for their future version within the information system**
- Need of new vocabularies > these too will have to be integrated later on (need of more flexibility)
- Since May 2021 the RDF/RiC-O files are available at https://github.com/ArchivesNationalesFR/Referentiels
- They will soon be available on the French government open data platform

Using RiC-O datasets to address the needs of several research projects

- **ALEGORIA** : https://www.alegoria-project.fr/

  - linking the places authority data with those of the IGN

  - several finding aids enriched and indexed, then converted to RDF/RiC-O, describing collections of aerial photos down to the item level

  - the metadata of two other institutions have also been converted to RiC-O/RDF

  - the whole will be available through a RDF store and interface soon
- A small ongoing project within the ANF, with the Departement Justice and Intérieur: a test on a small set, representative of a huge amount of **files created by the Direction générale de la sûreté nationale between 1870 and 1940, about people under surveillance** (immigrants, political opponents, etc.). Target: show both the documentary network between those records, the reuse of records by the police administration through time, and help better know the persons concerned.
- **ORESM** : https://oresm.hypotheses.org/a-propos/presentation-du-projet: research project carried out by the Bibliothèque interuniversitaire de la Sorbonne and the Laboratoire de Médiévistique occidentale, aiming to reconstruct the history of the archives of the University of Paris in the Middle Ages and gather data on students and staff of the university.

Preparing the future of the information system

Two main directions for now:

- as already said, working on the future data models and functionalities
(with a special care, not only to the authority data model, but also to the evolution of the EAD and EAC-CPF formats)
Building a model for the ANF, selecting what we need in RiC and RiC-O, and extending it if necessary (using other ontologies if necessary, e.g. PREMIS).
Need to update the model through time.

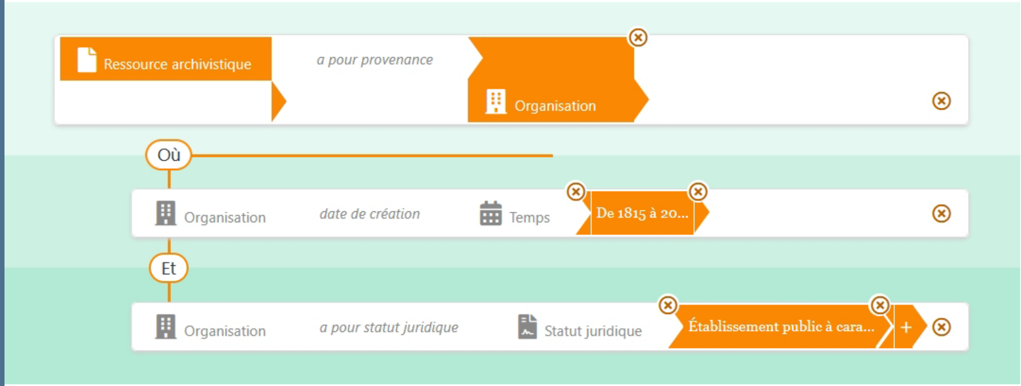
- working on research and visualisation tools for the RDF sets

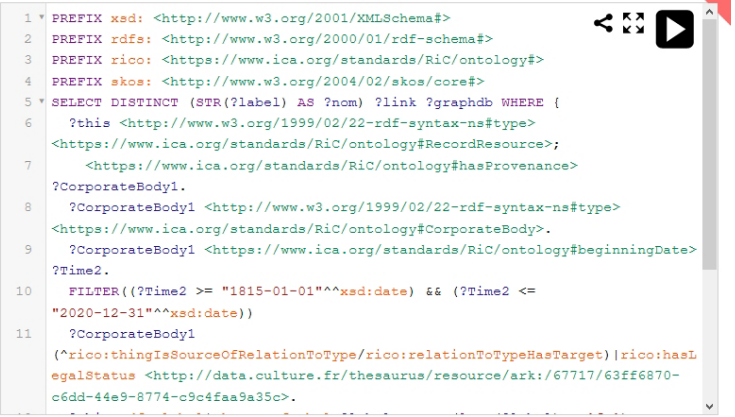
Querying RDF knowledge graphs more easily

Two barriers have to be removed:
- an end user will most often know nothing of SPARQL language
- an end user will most often know nothing of the domain model (here, RiC-O)

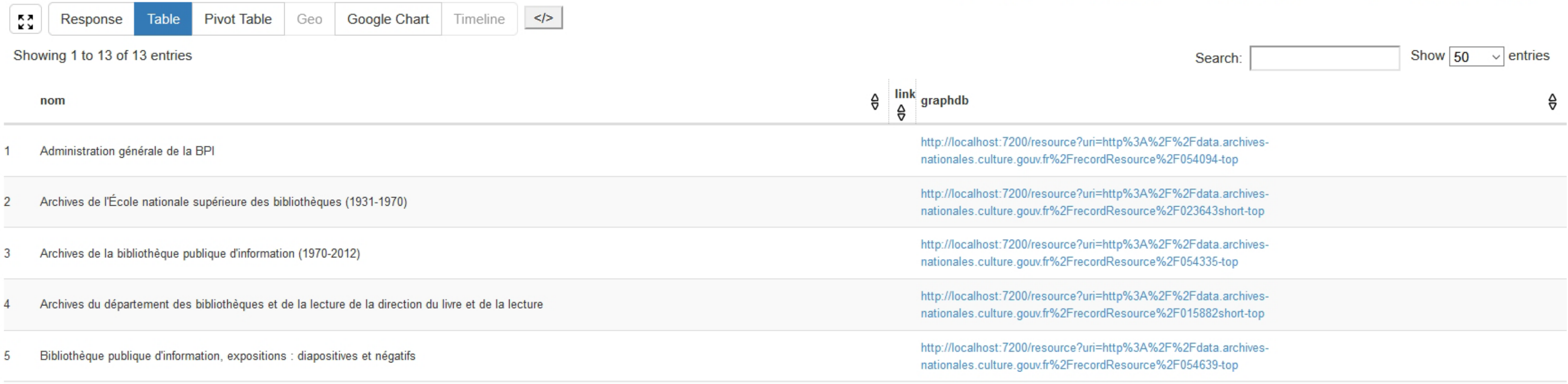
- Project aiming to improve Sparnatural (http://sparnatural.eu/), a visual SPARQL query builder, and making the results available for any institution or project
- Four partners involved : the ANF, the BnF, the ministry of Culture and Sparna private company
- Project started this summer, should end by February 2022
- A demonstrator (SPARQL endpoint + web interface integrating Sparnatural) to query a significant amount of the ANF RDF datasets, will be online at the end of the project.
- The project will include representatives of end users (through workshops)

Sparnatural: to build a SPARQL query on RiC-O datasets





```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rico: <https://www.ica.org/standards/RiC/ontology#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT (STR(?label) AS ?nom) ?link ?graphdb WHERE {
  ?this <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://www.ica.org/standards/RiC/ontology#RecordResource>;
    <https://www.ica.org/standards/RiC/ontology#hasProvenance> ?CorporateBody1.
  ?CorporateBody1 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://www.ica.org/standards/RiC/ontology#CorporateBody>.
  ?CorporateBody1 <https://www.ica.org/standards/RiC/ontology#beginningDate> ?Time2.
  FILTER((?Time2 >= "1815-01-01"^^xsd:date) && (?Time2 <= "2020-12-31"^^xsd:date))
  ?CorporateBody1 (^rico:thingIsSourceOfRelationToType/rico:relationToTypeHasTarget)|rico:hasLegalStatus <http://data.culture.fr/thesaurus/resource/ark:/67717/63ff6870-c6dd-44e9-8774-c9c4faa9a35c>.
```



Showing 1 to 13 of 13 entries

Search: Show 50 entries

	nom	link	graphdb
1	Administration générale de la BPI		http://localhost:7200/resource?uri=http%3A%2F%2Fdata.archives-nationales.culture.gouv.fr%2FrecordResource%2F054094-top
2	Archives de l'École nationale supérieure des bibliothèques (1931-1970)		http://localhost:7200/resource?uri=http%3A%2F%2Fdata.archives-nationales.culture.gouv.fr%2FrecordResource%2F023643short-top
3	Archives de la bibliothèque publique d'information (1970-2012)		http://localhost:7200/resource?uri=http%3A%2F%2Fdata.archives-nationales.culture.gouv.fr%2FrecordResource%2F054335-top
4	Archives du département des bibliothèques et de la lecture de la direction du livre et de la lecture		http://localhost:7200/resource?uri=http%3A%2F%2Fdata.archives-nationales.culture.gouv.fr%2FrecordResource%2F015882short-top
5	Bibliothèque publique d'information, expositions : diapositives et négatifs		http://localhost:7200/resource?uri=http%3A%2F%2Fdata.archives-nationales.culture.gouv.fr%2FrecordResource%2F054639-top

Leads that we have begun to investigate: using RDF to better assess some quality issues

Examples:

- Issue: the French government and administration entities change through time; our EAC-CPF records contain, among other ones, chronological and hierarchical relations (high temporal granularity), but nothing is available yet for displaying them

Target: build an interface that displays these changes, articulated with a timeline, and helps our colleagues to check their work, and users to explore the dataset

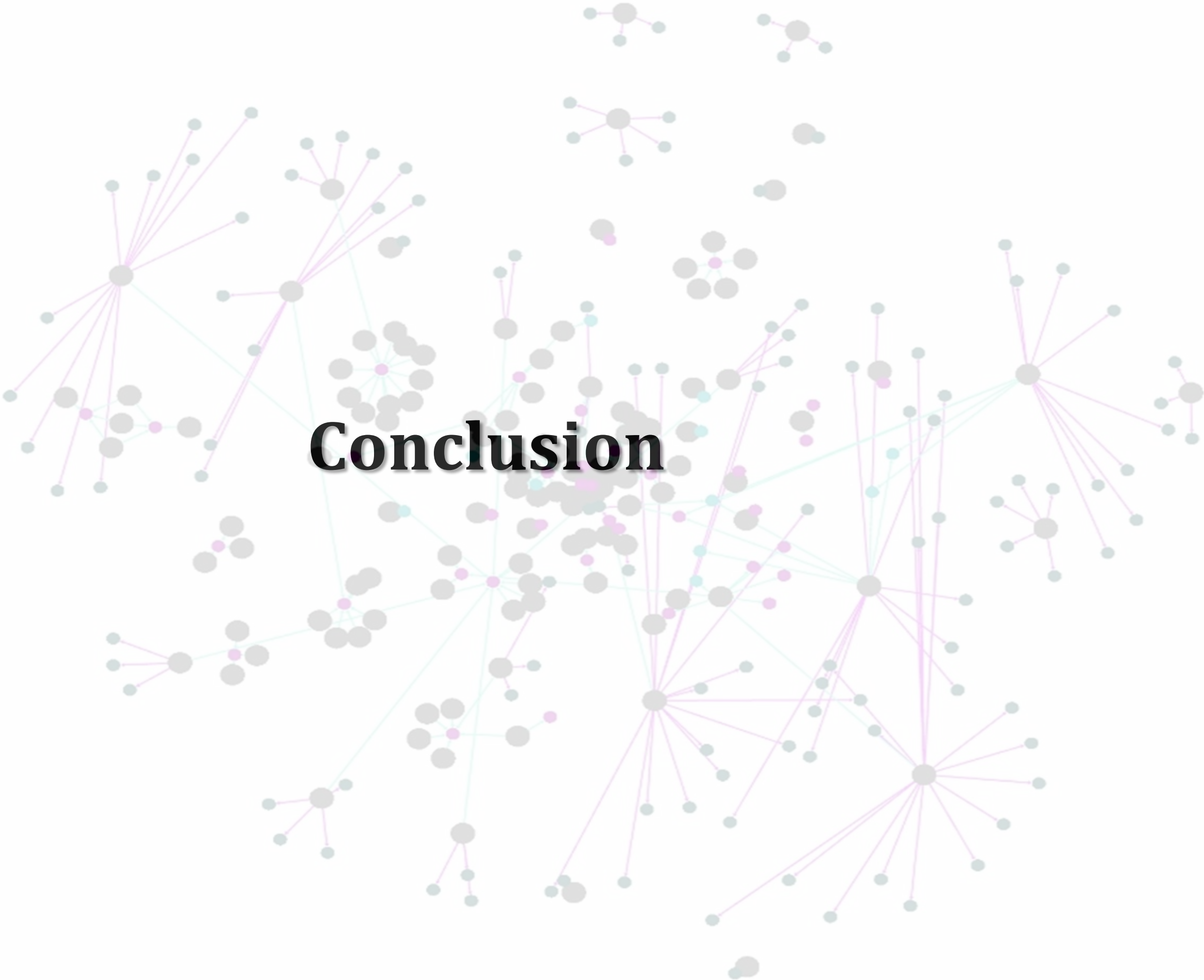
- Issue: through time, the same records sets may have been described several times in multiple finding aids.

Target: list, using a SPARQL query, the description units of finding aids that describe the same record resources, in order to (at least) link them, or even merge them.

Last but not least….

Defining and applying **a global, institutional, strategy for improving the quality of the metadata** :

- principles and objectives (both realistic and ambitious, aiming to adress the needs and expectations of the end users)
- going further and assessing the existing files more accurately
- defining and settling the methods and resources needed to reach the objectives, including human resources and tools



Conclusion

General outcomes

- A lot of work done, from first qualitative experiments to a far larger scale, whose results are encouraging and stimulating.
  These results also include some tools and methods.

- **The ANF projects also have been, and are, an opportunity and a way to test and enhance the reference standard**, which is still a draft.

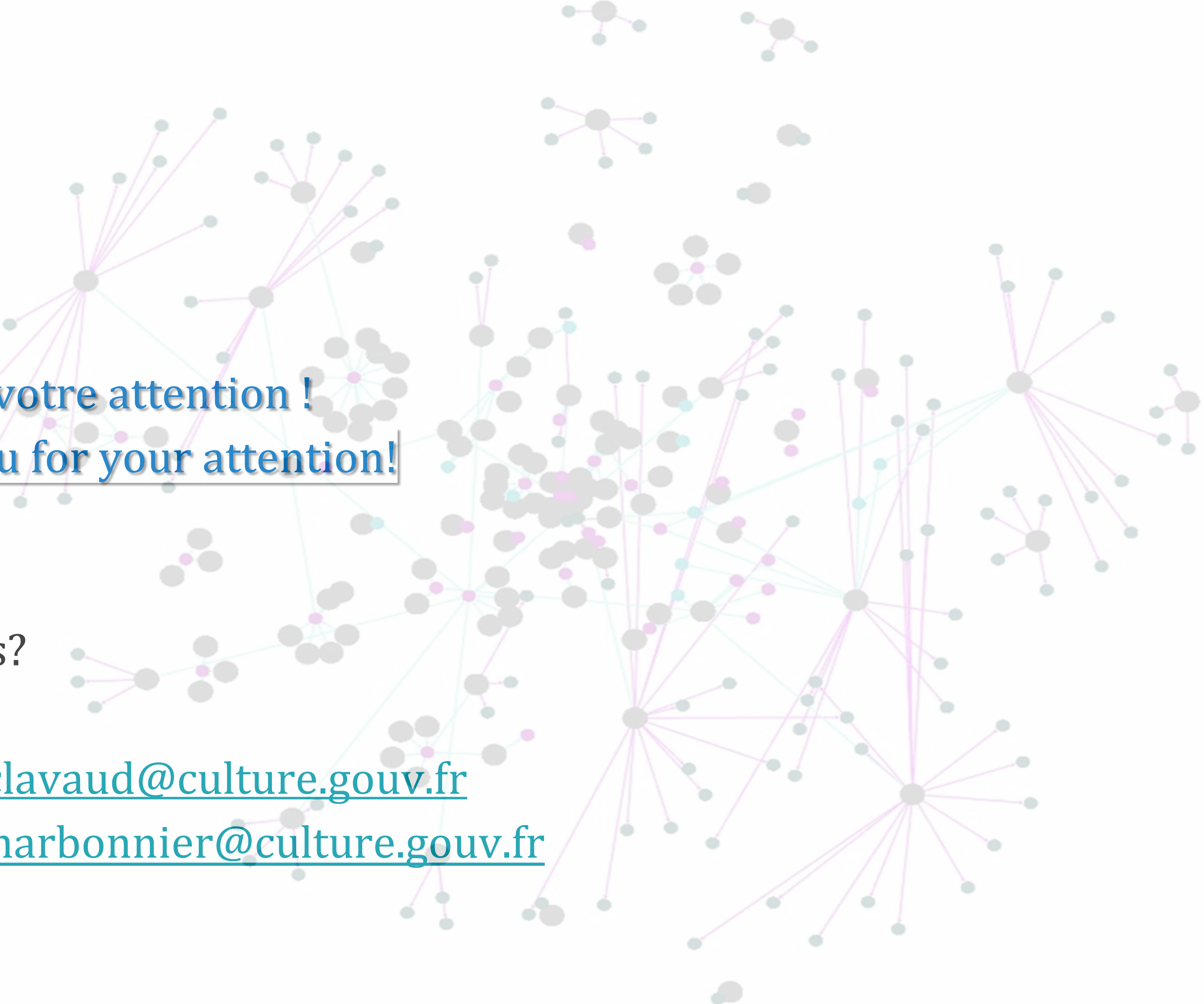
- **A lot of work still to be done, a lot of questions to be answered.**

- A lot of tasks can only be done through **collaborative programs**, without losing our domain perspective.
  Also need to ensure a technological watch.

Moving to Linked Open Data, one aspect of a global strategy

- Moving to a graph of linked entities is far from being only a technical issue; it fits within a **global issue: enhancing the quality and usability of archival metadata, moving towards FAIR data.**
  End users must be involved.

- Need to include this work in a **global trajectory.**
  Strategic plan for 2021-2025 is being designed just now, it takes into account these prospects.
- The ANF strategy must be coordinated with the French archival network and the France Archives portal.
- In the longer term, the ANF are interested in building international authority data for archives (e.g. about functions).



Merci de votre attention !

Thank you for your attention!

Questions?

Email:

florence.clavaud@culture.gouv.fr

pauline.charbonnier@culture.gouv.fr