MINISTÈRE DE LA CULTURE
Liberté
Égalité
Fraternité

ARCHIVES NATIONALES

RiC at the Archives nationales de France (AnF): an already old history, its milestones, first results and prospects

Florence Clavaud, head of the Lab at the AnF
Member of ICA/EGAD, lead of RiC-O development team
Email: florence.clavaud@culture.gouv.fr

Records in Contexts - And now? Informal exchange on the implementation of RiC,
organized by the KLA Committee on Information Systems (AFIS). Landesarchiv Berlin, April 23, 2024

A few elements of context

- The AnF: for now 380 linear kms of analogue records; about 70 To of born digital records (more key figures here)

- An old institution, whose current staff inherits the production of centuries of work (by archivists, historians, or persons who had both occupations, who created successive layers of findings aids).
  Our **metadata sets have their own history**

- **Several metadata sets in various formats**. A lot of databases, and at the core of the system, about 32000 EAD finding aids and 16000 EAC-CPF authority records on archival creators, plus about 20 vocabularies.
  So a huge amount of heterogeneous data.

- A complex, quite old now, siloted information system:
  - **metadata splitted in various silos, sometimes redundant, probably sometimes inconsistent, with difficulties to assess this data**
  - finding aids almost being 32000 microsilos - except that archival creators are being described, which creates **links between the finding aids, and the creators or accumulators of the records described**

- Metadata searchable through the online reading room (+ several other interfaces depending on the nature of the data), but not accessible by machines, apart from an OAI-PMH repository or some datasets on the open data platform of the ministère de la Culture

About the 15484 EAC-CPF authority records encoded in EAC-CPF and validated (in March 2022)

Relations specified between the authority records	**Number of relations**
Hierarchical Relations	9833
Temporal Relations	4434
Associative Relations	6455
Family Relations	529
Total	**21252**
Identity Relations (with external authority records)	**2345**
Relations with finding aids	**34021**

The relation that exists between two entities is counted once

RiC as a general framework to achieve a new (r)evolution

- This current state results from **decades of efforts** (the first digital (r)evolution)

- From 2013 (when its second new building was open to the public) a strategic plan has included programs that aim to improve the quality and accessibility of the metadata. This include calling for volunteers to index digitized items (crowdsourcing), systematically describing archival creators, or NER in the EAD files

- This is where RiC comes in.

  Our vision: **RiC as a global framework reference to move from silos to a data oriented architecture, rationalize the description of records, improve its completeness, precision and quality, publish this metadata as FAIR, Linked Open Data, and thus provide a better service to the public**

- **This is far from where we are**, and should be considered **a second (r)evolution**.

Question: How can we make this vision both a more precise and shared one, and move forward?

The first questions to be answered

- **Is it possible to move from our existing metadata to RiC based graphs**?
  - A qualitative proof of concept, PIAAF
- If so, **how can we move to the scale of our huge amount of data**?
  - Developing tools, workflows, and more (ongoing)
- **How can we value a RiC-O based graph**, particularly for our public (general public and researchers)? What benefits to the public?
  - A web application interface (ongoing)

- Some of our strengths and weaknesses:
  - a very active involvement in EGAD activities from the beginning, particularly in developing RiC-O
  - a very small but very motivated team, with limited financial means (but continuous financial support from the French Ministère de la Culture)
  - connections with a lot of partners outside
  - work to be done outside the information system

The PIAAF proof of concept

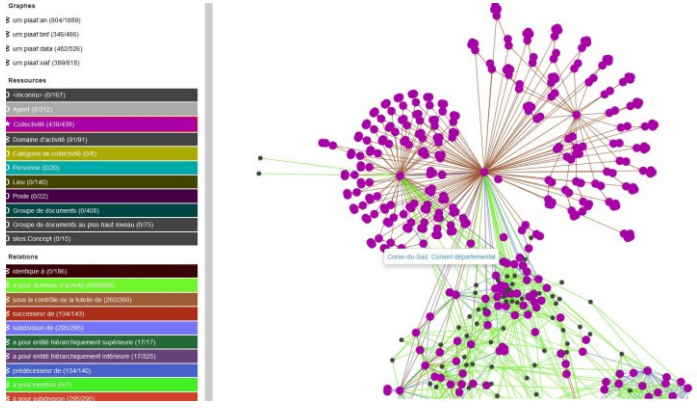
- The **first RiC-based, qualitative, proof of concept in the world.**
- Data semantized by the AnF, interface developed by a French private company (Logilab).
- Released in **February 2018**: https://piaaf.demo.logilab.fr/
- Proved:

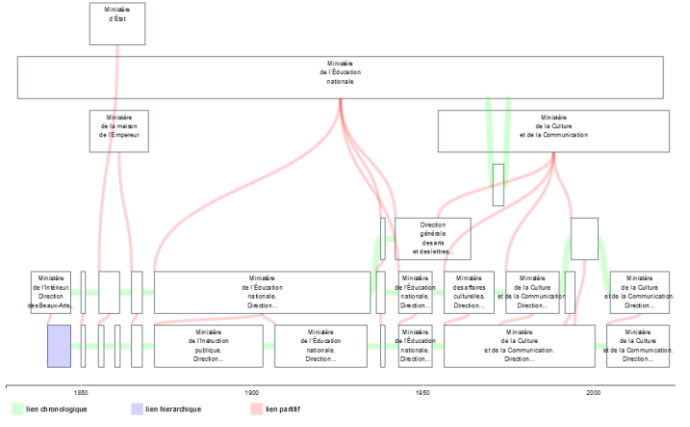
  - that 'legacy' metadata can be converted to RiC-O/RDF sets without losing anything of their content

  A lot of manual work to check the quality of 3 small sets (about 280 EAC-CPF authority records and 40 EAD finding aids, and a shared vocabulary on activity and domain types)

  - that sets of RDF data can be connected to each other, using for example the ISNI number of agents; and this opens new perspectives for enriching each dataset

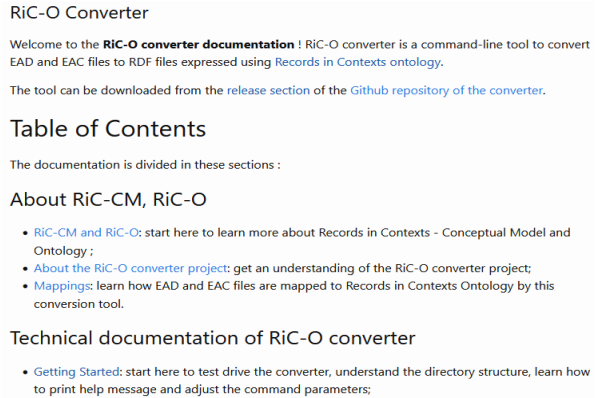
  - that querying such data using SPARQL enables to ask questions that could never be asked before; that the resulting graphs can be displayed and browsed as such, revealing the relations existing between the XML source files





Other findings:
- some work on data visualization, e.g. showing the chart of corporate bodies through time, or displaying the description of a n-ary relation
- early version of RiC-O extensively tested
- See also this presentation (in English): https://enc.hal.science/hal-03958855

Moving to a much larger scale: developing RiC-O Converter



RiC-O Converter

Welcome to the **RiC-O converter documentation** ! RiC-O converter is a command-line tool to convert EAD and EAC files to RDF files expressed using Records in Contexts ontology.

The tool can be downloaded from the release section of the Github repository of the converter.

Table of Contents

The documentation is divided in these sections :

About RiC-CM, RiC-O

- RiC-CM and RiC-O: start here to learn more about Records in Contexts - Conceptual Model and Ontology ;
- About the RiC-O converter project: get an understanding of the RiC-O converter project;
- Mappings: learn how EAD and EAC files are mapped to Records in Contexts Ontology by this conversion tool.

Technical documentation of RiC-O converter

- Getting Started: start here to test drive the converter, understand the directory structure, learn how to print help message and adjust the command parameters;

- An open source, powerful and fast, reliable, easy to install, configurable and adaptable software, to convert EAD 2002 and EAC-CPF files into RDF/RiC-O datasets
- Developed with Sparna private company; based on **mappings and precise unit tests prepared by the AnF and Sparna**, which are also available in the source code.
- Now **documented in English and French**.
- v1.0 released in April 2020
- **v2.0 released in October 2023 (produces data compliant with RiC-O 0.2)**
- **v3.0 will be released in 2024** (among other developments, will allow the production of datasets compliant with RiC-O 1.0); a v4 could take the descriptions of born digital records into account
- See also this article: https://doi.org/10.1145/3583592 (published in August 2023, in the JOCCH, vol. 16, issue 3)

https://github.com/ArchivesNationalesFR/rico-converter

Findings and open questions:
- patterns and methods for generating URIs
- mappings between EAD 2002 and EAC-CPF and RiC-O
- quality issues in the source metadata revealed or better known, as well as their consequences
    - among the main issues: the route from ISAD(G) description units to Records, Record Parts and Record Sets
- need to work on application profiles (SHACL shapes) to better control the quality of the output
- what should ISAD(G) description rules become in a world of graphs?

Moving to a much larger scale: working on the authority records and vocabularies

Référentiels des Archives nationales de France

Les référentiels des Archives nationales de France | the Archives nationales de France authority data and vocabularies

Téléchargement

Télécharger la dernière release (la version 1.1, juin 2022): https://github.com/ArchivesNationalesFR/Referentiels/releases/tag/1.1

Description

Formats :

- XML/RDF pour l'ensemble des référentiels. Principaux modèles utilisés : SKOS et Records in Contexts - Ontology (RiC-O) (dans sa dernière version officielle en date, la v0.2 de février 2021)
- XML/EAC-CPF pour les notices du référentiel des producteurs
- CSV (encodage UTF-8; séparateur de champs : virgule ; valeurs des colonnes encadrées par des guillemets ("")) pour les concepts et les lieux. Ces fichiers CSV ont été produits à partir des versions RDF. Pour les agents, les fichiers CSV fournis sont des listes, qui ne contiennent pas toute la substance des notices RDF des agents.

En cours de construction | a work in progress...

Licence | License : Ces métadonnées étant des informations publiques, l'usager dispose d'un droit non exclusif et gratuit de libre « réutilisation » à des fins commerciales ou non, dans le monde entier et pour une durée illimitée. Il doit accompagner chaque rediffusion des informations de l'indication précise de l'origine des métadonnées : « Archives nationales (France) », date de ces métadonnées (mai 2022), nom du référentiel (fourni dans le fichier Excel qui en donne la liste). Voir à ce sujet la page : https://www.archives-nationales.culture.gouv.fr/fr/web/guest/reutilisation-des-donnees-publiques.

La liste fournie (fichier Liste-referentiels.xlsx) donne diverses informations sur chacun de ces référentiels.

Avertissements

Les référentiels sont **en construction**, ce qui veut dire qu'ils sont en règle générale très incomplets : il y manque la description de nombreuses entités, et les descriptions fournies sont parfois pauvres. Ils sont cependant utilisés, dans un format propre au SI des Archives nationales, pour indexer la description des archives que l'institution conserve, ou pour

https://github.com/ArchivesNationalesFR/Referentiels

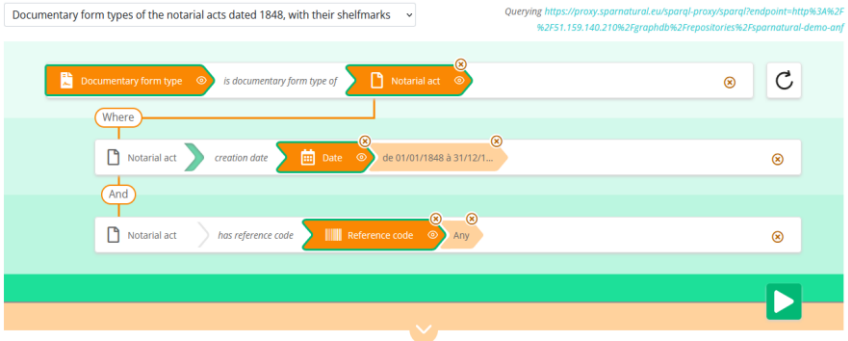
Authority records on agents and places, and vocabularies created and used by the AnF in their IS, to describe (index) the records kept:
- result from decades of effort
- **none of the files are, however, yet directly accessible to end user**s (except the EAC-CPF records about the archival creators, which can be displayed in HTML in the virtual reading room)
  - this data set **should play a key role in the future**: it describes entities that are contexts of the records kept, and should be entry points for end users
- **First RDF release, based on RiC-O 0.2 and SKOS, published on GitHub in 2022**; EAC-CPF files are being semantized using RiC-O Converter, other datasets produced using a suite of dedicated scripts
- Also available in CSV (generated from the RDF)
- **New releases soon** (including moving to RiC-O 1.0 of course)

Findings and open questions:
- Data that are continuously being enriched as part of various projects, some of which use semantic technologies: RDF as an ecosystem
- Need of a reliable workflow
- **Need of a web search and consultation interface in 2024, and of an API,** in order to make this dataset accessible to the public, to archivists, and to machines; hopefully this project will start in 2024
- much more data in the various silos of the AnF could be considered authoritative data

Publishing and valuing a large knowledge graph

A demonstrator for searching and exploring a large knowledge graph:
- first version published online in **June 2022**
- **allows the general public to search a third of the metadata produced by the AnF to describe the archives of Paris notaries from the 15th century to nowadays**; i.e. approximately **60 million triples currently**, conforming to RiC-O 0.2 (slightly extended)
- Search interface built using the Sparnatural open source visual query editor, which allows **exploring the knowledge graph without knowing SPARQL or RiC-O**

And **to ask questions that cannot be asked in the current AnF IS**
- Available in French and English, like its documentation; dataset also available on GitHub: https://github.com/ArchivesNationalesFR/Sparnatural_prototype_data



https://sparna-git.github.io/sparnatural-demonstrateur-an/

Among the findings:
- though all RiC-O is not used, an extension of RiC-O is definitely necessary for the AnF, as well as building sets of inference rules (so handling distinct named graphs)
- Redundancies revealed at the heart of the graph (as concerns the descriptions of some records)
- Using such an intuitive, interactive exploration interface requires a greater intellectual commitment (than using old search forms)

The project will continue: making the data and interface compliant with RiC-O 1.0, developing a new version of the query editor with new possibilities, and implementing it in the demonstrator, integrating new versions of the authority data and vocabularies, and perhaps extending the scope to other metadata

Prospects

Already a variety of new perspectives concretely explored and new possibilities; an iterative process

There is a lot to do apart from working on RiC-O Converter and the demonstrator, and **this first concerns the datasets themselves**. The following is in our TODO lists (or has already started):

- Through the first project previously presented, working on reducing the redundancy at the heart of the data and backpropagate the outcomes in the source EAD files
- Producing one or more SHACL application profiles aiming to better control the quality of the RDF output
- Working continuously on the authority records and vocabularies
  - in particular, in 2024, a major work in progress on the nomenclature of the roads of Paris, and one on the forms, types of acts and states of documents, whose current SKOS serialization will be enriched using RiC-O properties
  - plus other more specific projects aiming to generate additional authority records on agents directly from the databases or finding aids
- Contributing to research projects like ORESM (which should be an opportunity to both use and enrich our datasets)
  - developing and implementing one or more extensions of RiC-O in order to describe some records more accurately, or to store uncertainty (in equivalences or as concerns the relations), and the evidence of assertions
- Using our knowledge graphs and AI to develop a Named Entity Recognition and Linking prototype

As a more general conclusion (to this short presentation…)

- Lessons learned:
  - **You do not need to wait for your system to be 'ready for RiC', nor to have a lot of resources, to work on your data and use RiC**
  - This is basically about **metadata processing, curating and publishing (moving to FAIR data)**, and thus it is a step by step, iterative and long-term process
  - A wide range of benefits can be anticipated from the very beginning, from higher quality, rationalized data, to a far better service to the public
    and this may only be the beginning of what we can expect
  - RiC is a global framework, which can embed and 'bind' much more than the classic metadata we are used to thinking when we think of descriptive metadata
  - An opportunity to **work collaboratively** - in fact it is necessary
  - Also an opportunity to make the archival community closer to the researchers in history, humanities and social science
  - And of course, **an opportunity to contribute to the development of RiC ;-)**

Thank you for your attention!
Any questions or comments?

(do not hesitate to contact me: florence.clavaud@culture.gouv.fr)