RiC-O Converter: a Software to Convert EAC-CPF and EAD 2002 XML files to RDF Datasets Conforming to Records in Contexts Ontology

Thomas Francart, Florence Clavaud, Pauline Charbonnier

RiC-O Converter: a Software to Convert EAC-CPF and EAD 2002 XML Files to RDF Datasets Conforming to Records in Contexts Ontology*

Thomas Francart[1], Florence Clavaud[2], and Pauline Charbonnier[2]

[1] Sparna, Tours, France
[2] Archives nationales, Pierrefitte-sur-Seine, France
thomas.francart@sparna.fr, florence.clavaud@culture.gouv.fr, pauline.charbonnier@culture.gouv.fr

Abstract. RiC-O Converter is an open-source command-line tool to convert EAD finding aids and EAC-CPF authority records to RDF files conforming to Records in Contexts ontology, in a robust manner. It was developed for the Archives nationales of France (ANF) but is aimed to be reused by other archival institutions, and to this aim is fully documented in English. It is based on XSLT stylesheets that take into account the variability of EAD content. It enabled the ANF to convert 15000 EAC-CPF files and 29000 EAD files into an homogeneous knowledge graph. Such a graph opens new perspectives for navigating and linking from/to archival metadata.

Keywords: Records in Contexts (RiC), RiC Ontology (RiC-O), RDF, XML EAD, XML EAC-CPF, open source software.

1 Introduction

RiC-O Converter is an open-source command-line tool to convert EAD finding aids and EAC-CPF authority records to RDF files conforming to ICA Records in Contexts ontology.[1] The tool, ordered by the Archives nationales of France (ANF), was developed by Sparna, a French company specialized in semantic Web and knowledge graphs engineering. The Department of digital innovation of the French Ministry of Culture sponsored and funded the project according to the semantic roadmap the ministry is conducting. The tool was released on GitHub in April 2020.[2]

[1] ICA Records in Contexts Ontology (RiC-O) [1] is presented in another article authored by Florence Clavaud and Tobias Wildi.

[2] RiC-O Converter source code: https://github.com/ArchivesNationalesFR/rico-converter (last accessed 2021/07/03).

2 Project History

Many archival institutions and projects (like portals such as Archives Portal Europe and FranceArchives) around the world use XML/EAD and XML/EAC-CPF files to describe their collections and the agents related to them. Based on ISAD(G) standard, XML/EAD (EAD) [2] facilitated its adoption and diffusion. EAD either is the production and storage format of finding aids or is the output format of data bases. It preserves finding aids and is an exchange format with external applications. Based on ISAAR(CPF) standard, XML/EAC-CPF (EAC) [3] is quite often used to describe authorities such as corporate bodies, persons and families that created or accumulated the fonds held by archival institutions.

RiC-O Converter project is based on the statement that transforming EAD and EAC files to RDF, thus creating knowledge graphs about archives and their contextual entities, results in an homogeneous and interoperable data structure, that is compliant with the FAIR principles,[3] and opens new perspectives related to querying, browsing, reusing, publishing and linking from / to archival metadata.

The ANF[4] have been interested in entity-relationships models and graph technologies since 2013, one of the reasons being that they already have authored a significant, and growing, number of authority records that were linked to each other and to the descriptions of the archives themselves. This in essence constitutes a very dense oriented graph, whose relations are not really displayed, and cannot be queried and processed in the ANF current information system [4, 5]. The ANF also wanted to connect these metadata with other metadata sets created by other institutions. Linked Data technologies thus seemed to be a possible solution to meet these needs. RiC Ontology, an OWL domain ontology for archives, was at last available; it is also based on a recent entity-relationship conceptual model [6]; it is fully documented and rich. Which made it possible for the ANF to produce RDF datasets.

The ANF first built a qualitative proof of concept (PIAAF)[5], to show that converting existing archival metadata to RDF datasets conforming to RiC-O was possible, to interconnect datasets from different institutions, and to visualize and explore them in a new way. But the PIAAF prototype did not include a large quantity of metadata nor took into account the variety of their structure and content. Therefore, the ANF needed to move from this qualitative proof of concept to a large-scale project. Indeed, the ANF hold a significant amount of metadata to be converted, which implied to develop a reliable, efficient, configurable, tool. The tool was designed to process finding aids and authority records only, even if the ANF also hold several controlled vocabularies. Building RDF/RiC-O vocabularies is a different process and is still a work in progress

[3] The FAIR Guiding Principles for scientific data management and stewardship (2016) are recommendations to improve "the Findability, the Accessibility, the Interoperability and the Reuse of digital assets", around which a community and various initiatives have developed. See the website: https://www.go-fair.org/ (last accessed 2021/08/23).

[4] Homepage of the ANF website: http://www.archives-nationales.culture.gouv.fr/, last accessed 2021/08/23.

[5] Homepage of the PIAAF project website: https://piaaf.demo.logilab.fr/, last accessed 2021/07/03.

in the institution, since the internal data structure of these controlled vocabularies, that currently conform to a very poor, locally defined, model, should change soon.[6]

The conversion tool was required to be **industrial (**meaning: *performant* and capable of processing tens of thousands of input files in a reasonable amount of time); *tested* (to guarantee the coverage of all possible situations encountered in source files); *verbose* (to produce log files to follow its execution), **easy to install** (so it can run on a typical desktop machine), **configurable** and **adaptable** (to suit the ANF needs as well as different needs of potential reusers), well **documented**.

3 Design and Main Characteristics

The proposed solution relies on XSLT stylesheets, encapsulated in a Java script. The Java wrapping of the XSLT ensures a convenient command-line interface, the proper sequencing of the conversion steps, and portability to all operating systems.

The stylesheets convert EAD and EAC in RDF/XML containing instances of RiC-O classes and properties. They live in a separate directory from the java command itself, ensuring that modifications can easily be made in conversion logic without the need to recompile the tool.

The development methodology relied on unit tests to cover all possible situations that could be found in input files. Each of the 90 unit tests is specified in an input EAD or EAC file with a corresponding expected RDF/XML file.[7] When run, the output of the converter is compared to the expected file, and, if differences are found, the test fails. The tests can be run directly from the command line, so that any user can verify the tests, and add its own, if the stylesheets are modified. The tests ensure no regressions are introduced when the software evolves.

Running the tool is as simple as running a bat or sh script. The script asks for the action to be executed (EAC or EAD file conversion) and the option properties file to use.

The EAC to RIC-O conversion process is summarized in the following diagram:

---

[6] About the context and the ANF ongoing projects, see presentations from the study day dated January 28, 2020, on Les métadonnées archivistiques en transition vers des graphes de données: https://labarchiv.hypotheses.org/1495, especially https://labarchiv.hypotheses.org/files/2020/02/20200128_3_RiCauxAN_EnjeuxPremieresRealisations.pdf (in French). See also a more recent presentation by Florence Clavaud, "Implementing ICA Records in Contexts-Ontology at the National Archives of France: first assessment and prospects", for the Study Day on The Semantic Web and Cultural Heritage: From Data Convergence to Knowledge Crossing (Lille, France, February 3, 2021) - slides in English and audio recording in French: https://geriico.univ-lille.fr/detail-event/le-web-semantique-et-le-patrimoine-culturel-de-la-convergence-des-donnees-au-croisement-des-connai/ (last accessed 2021/07/03).

[7] The unit tests are available at https://github.com/ArchivesNationalesFR/rico-converter/tree/master/ricoconverter/ricoconverter-convert/src/test/resources (in two subfolders named eac2rico and ead2rico). One input/output file encodes more than one test.

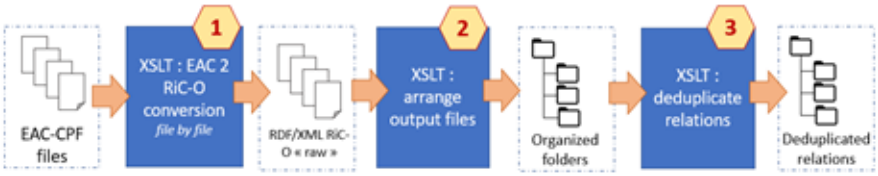

Fig. 1. EAC-CPF to RiC-O conversion process

1. Input EAC files are converted into RiC-O RDF/XML. Each input file yields a corresponding RDF/XML file. An option allows to stop processing here to examine the raw output of the conversion.
2. The content of the raw RDF/XML files is reorganized to split the output in folders corresponding to agents, places, and relations. Relations are grouped into large files, each corresponding to a high-level relation class in RiC-O: Agent Hierarchical Relations, Agent Origination Relations, Agent Temporal Relations, Agent To Agent Relations, Family Relations, Membership Relations, Work Relations.
3. The relations are deduplicated to remove those that appear more than once. As the original relation is expressed in the source files for both related entities, the same relation expressed in RiC-O was generated twice in step 1.

The EAD to RiC-O conversion process is summarized in the following diagram:

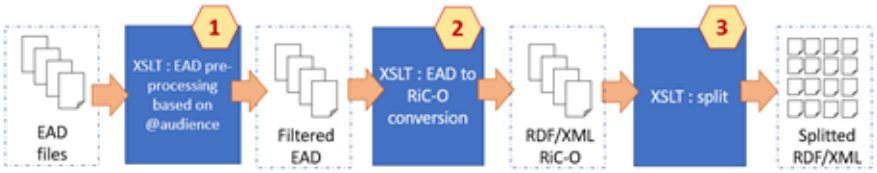

Fig. 2. EAD to RiC-O conversion process

1. Input EAD files are filtered according to the "@audience" EAD attribute, so that non-public files or components of files can be excluded from the conversion process.
2. The filtered EAD files are then converted into RiC-O RDF/XML. Each input file yields a corresponding RDF/XML file.
3. If requested, the output files can be splitted into smaller files, with the top Record Resource in one file, and each "branch" of the finding aid in a separate file.

It should be noted that the conversion step to RiC-O RDF/XML takes the assumption that each "c" element in the input EAD file has an "@id" XML attribute from which the corresponding Record Resource URI is derived.

The EAD conversion takes into account some variability of what can be found in EAD files, where the same element is allowed to contain different content. For example, a "physdesc" element with only text, or a "physdesc" element with mixed content including "dimensions" and "extent" children elements, or a "physdesc" element

with “extent” and “physfacet” children elements, containing a reference to a controlled vocabulary, result in different outputs, as shown in the corresponding unit test.[8]

The performance of the tool is very good: 15200 EAC files are processed in approximately 15 minutes, yielding 0.7 million triples. 29000 EAD files are processed in approximately 30 minutes, yielding approximately 155 million triples.

4 A Tool for the Archival Community

The software is fully documented in English to reach the international community. The project team produced mappings between EAC-CPF and EAD to RiC-O, which will be very useful to the archives community, especially institutions considering converting their metadata to RDF. The mappings are available in the documentation section[9] of the tool. The conversion process and how to customize the conversion are documented too. The source code is open and freely accessible on GitHub. The software is licensed under the terms of the CeCILL-B license.[10]

The converter was developed to address first the needs of the ANF but we kept in mind its potential uses for any other archival institution; therefore the code is easily configurable. Typically, the root URI for the URIs to be generated is an option that can be easily changed. However archival institutions would probably need to adapt the software to their own systems. For example, we did not take into account some elements in the conversion process because these elements are not used in the ANF EAD finding aids (“abstract”, “dao”, “bioghist”, etc.) or because they are not relevant in RDF (“frontmatter”, “titlepage”, etc.).[11]

5 Results and Prospects

The ANF have converted nearly 29000 finding aids and 15200 authority records using RiC-O Converter. We can convert them again when needed, for example when major updates occur in our metadata.

Quality issues appeared during development time (lack of precision or worse, bad use of EAD format). More generally speaking, improving the quality of archival metadata is a key issue for the ANF, and quality management and data governance also

---

[8] Physdesc unit test: https://github.com/ArchivesNationalesFR/rico-converter/tree/master/rico-converter/ricoconverter-convert/src/test/resources/ead2rico/_32_physdesc (last accessed 2021/07/03).

[9] EAD to RiC-O and EAC-CPF to RiC-O mappings can be found in https://github.com/ArchivesNationalesFR/rico-converter/tree/master/ricoconverter/ricoconverter-doc/src/main/resources (EAC_to_Ric-O_0.1_documentation.xlsx and EAD_to_Ric-O_0.1_documentation.xlsx files).

[10] https://github.com/ArchivesNationalesFR/rico-converter/blob/master/license.txt.

[11] On these aspects, and on all the topics presented in this article, more information is available (in French) in [7].

has to be enhanced. In a way, processing the RDF datasets generated can help assess this problem and solve it. Examples include: aligning the data on the agents to other RDF datasets e.g. those of the French national Library (BnF)[12] or of Wikidata[13] in order to enrich them, or (not investigated yet but an important need) linking (merging) distinct descriptions, authored through time in distinct finding aids, of the same archival resources.

Also related to quality would be the use of SHACL[14] rules to assess the conformance of the generated graph structure against some rules; these rules could be derived directly from the RiC-O ontology[15] (typically cardinality, domain and range check), or they can be hand-written to validate business oriented patterns in the graph. This is something to be done in the future.

In terms of challenges, RDF datasets resulting from the conversion are not published or searchable yet because of the lack of infrastructure in the ANF information system. The ANF do not have any triplestore available online. However, the ANF published dumps of the RDF datasets.[16] Besides, the data are already used in research projects such as ALEGORIA,[17] a project that aims at facilitating the promotion of iconographic institutional collections describing the French territory in various periods going from the interwar period to our days. A triplestore accessible through a SPARQL endpoint will soon be released. It will be connected to a web application demonstrating 3D immersive navigation through geolocalised photographs.

Moreover, the ANF should release a quite large-scale prototype by the beginning of 2022, including an easy-to-use, visual, SPARQL query interface.[18]

More generally speaking, RiC-O Converter needs to evolve for several reasons. RiC-O Converter is based on RiC-O 0.1 (dated December 2019); but Records in Contexts has evolved since that date: RiC-O 0.2 was released in February 2021, and introduces new components (like the Extent class) as well as updates (particularly as concerns the names of several object properties). The corresponding changes will be made in RiC-O Converter before the end of 2021.

---

[12] The BnF provides a web interface, including a SPARQL endpoint, for its RDF datasets, at https://data.bnf.fr/. The datasets can be downloaded from the following page: https://api.bnf.fr/dumps-de-databnffr (pages last accessed 2021/08/23).

[13] About the Wikidata well-known knowledge base, see https://www.wikidata.org/ (last accessed 2021/08/23).

[14] Shapes Constraint Language, https://www.w3.org/TR/shacl/ (last accessed 2021/07/03).

[15] Using for example SHACL Play! See https://shacl-play.sparna.fr/play/rules-catalog (last accessed 2021/07/03).

[16] A small subset of the EAD and EAC files of the ANF, and their RDF/RiC-O 0.2 version, is available in the RiC-O Git public repository: https://github.com/ICA-EGAD/RiC-O/tree/master/examples/examples_v0-2/NationalArchivesOfFrance. The ANF also have started to publish their authority records and vocabularies (RDF version, using mainly RiC-O 0.2 and SKOS, available at https://github.com/ArchivesNationalesFR/Referentiels).

[17] https://www.alegoria-project.fr/en (last accessed 2021/07/03).

[18] This interface will be built with Sparnatural; see https://github.com/sparna-git/Sparnatural (last accessed 2021/07/03).

RiC-O Converter does not convert XML/SEDA[19] files, used in French digital archives management systems to describe these digital archives. These files include technical and preservation metadata, whose definition is inspired by PREMIS data dictionary[20] which is more widely known in archives. Mapping SEDA to RiC-O and its transformation is a task to do in a future version of RiC-O Converter.

6 Conclusion

The transition from existing formats to novel graph-based and web-oriented conceptual models represents a challenge that can hinder the adoption of such new models. We especially think of FRBR and LRM in the library world, or CIDOC CRM for museums. By providing RiC-O Converter, a robust, adaptable and off-the-shelf tool to transition from EAD and EAC to RiC-O, the archival community aims at soothing and encouraging this transition, in order to make archival data part of the Web of data.

References

1. International Council on Archives (ICA) Records in Contexts-Ontology (RiC-O) latest official release: https://www.ica.org/standards/RiC/ontology, last accessed 2021/07/03.
2. Encoded Archival Description (EAD): https://www.loc.gov/ead/, last accessed 2021/07/03.
3. Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF) XML schema: https://eac.staatsbibliothek-berlin.de/, last accessed 2021/07/03.
4. Clavaud, F.: Building a knowledge base on archival creators at the National Archives of France: issues, methods, and prospects. In: Journal of Archival Organization, vol. 12, 1-2 (2015), pp. 118-142. Doi:10.1080/15332748.2015.1001642.
5. Clavaud, F.: Transformer les métadonnées des Archives nationales en graphe de données : enjeux et premières réalisations, in: Les Archives nationales, une refondation pour le XXIe siècle, La Gazette des Archives, n°254 (2019-2), pp. 59-88.
6. International Council on Archives (ICA): Records in Contexts-Conceptual model (RiC-CM) 0.2 (July 2021), https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf, last accessed 2021/08/23.
7. Francart, T., Charbonnier, P.: RiC-O Converter, un logiciel libre de conversion de métadonnées archivistiques (en EAD et EAC-CPF) en jeux de données conformes à RiC-O (2020/01/28), https://labarchiv.hypotheses.org/files/2020/02/20200128_4_RiCOConverter.pdf, last accessed 2021/07/03.

[19] The SEDA (‘standard d’échange de données pour l’archivage’) models the various transactions that may occur between the agents involved in digital archiving. This French standard conforms to ISO 20614:2017 (Information and documentation - Data exchange protocol for interoperability and preservation). The standard includes an XML schema. See https://francearchives.fr/seda/index.html (last accessed 2021/07/03).

[20] PREMIS (PREservation Metadata Implementation Strategies) is a data dictionary that was created in 2005, and is now expressed, among other formats, through an OWL ontology. It is hosted by the Library of Congress and maintained by the PREMIS Editorial Committee. See http://www.loc.gov/standards/premis/ (last accessed 2021/07/03).