A FAIR archive based on the CERIF model

Intro/Background

Electronic archives are, in many cases, created in systems that are aimed for specific activities without interoperability in mind. If CERIF is employed in relevant archive processes, a FAIR [1]compliant archive can be easier to achieve.

The article discusses how a CERIF based archive structure will manifest itself in the different information package stages of the OAIS model[2] (SIP, AIP, and DIP).

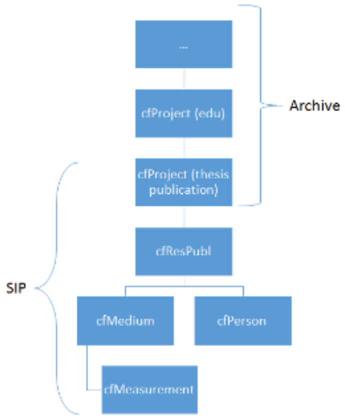
The CERIF model[3, 4] is used to represent research information and to transfer it between repositories. A special kind of repository is the electronic archive with some special requirements. We propose using CERIF as the main model for an archive and provides a real-life example from TheUniversity.

CERIF as archive

The main classification scheme is a tree of properly classed cfProject entities. This entities represent real processes, projects or activities that produce content or groupings of them in accordance with current archival recommendations that documents should be described from a process perspective [5, 6]. Naturally, the cfProject entity can also be archive file within an archive creator in a more classical sense, for example a department or other organizational unit.

Subtrees can consist of independent sub-archives and the entire tree can be joined to other trees to form a bigger tree in the OAIS ingest process[2]. For example, reporting the content of an institutional archive to a national archive would be achieved by copying the local tree with the relevant local information attached to the global tree. In the same way branches of the archival tree could be managed by a sub-unit in a quite independent way.

Most of the archived objects are to be represented by CERIF result entities. The metadata can be represented by CERIF entities attached to these, for example cfMedium to represent files.



Figur 1. The process archive of the student project work in CERIF terms. The main process is "education processes" (edu), sub-process "thesis publication". The published result is the PDF-file (cfMedium) with associated cfMeasurement containing embargo (premis) information

Master data like the organisation structure, lists of person, and infrastructure can be kept and managed at the most convenient and effective level enabling information sharing by several archival packages. This will also work towards fulfilling the GDPR criteria on minimizing personal information.

Electronic archives must be able to record events affecting the data or the structure of the archive. In CERIF context, this can be done attaching at cfMeasurement where Premis[7] XML data is stored. The object identifiers in the Premis model are equated to the CERIF entity identifiers, making the reference to them transparent.

Describing the data and structure of an archive in terms of CERIF allows exporting and, to some extent, importing data even when the actual structure is not a CERIF database. The use of CERIF in archives brings to this area the interoperability that repositories have had for a while (FAIR!).

The scope of sub-archives

Archives are built as close to the activity that produces the material as possible. This closeness can be given by the organization of the business in terms of business areas or processes or in terms of physical distance. A mechanism to simplify the delivery to a common archive of the information contained in several sub-archives is desirable.

Each sub-archive "knows" about the structure of underlying archival units in the form of a tree of cfProject entities classified as "archival units". The top node could hold a list of external archives to witch the unit belongs.

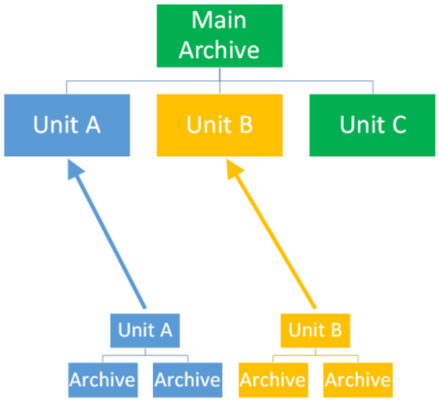


Figure 2. The archive process should be close to the organisational unit producing the actual content. If the involved units transfers the archive information in CERIF the main archive will have a trivial task organising the information.

Identifiers

Every CERIF entity has an identifier. Although it is not mandatory with globally unique identifiers, UUIDs are often used. In our case globally unique identifiers, and explicitly stated in the XML, are crucial for transferability. When objects in an existing system have a natural identifier, a version 5 UUID can be constructed based on the system’s URL. In that way the UUID can be recalculated from the internal identifiers whenever needed. In many cases random version 4 UUIDs will work just fine. The practice at TheUniversity is to use version 5 UUID identifiers constructed from, where applicable, local identifiers combined with a UUID of the domain name uni.edu as name space.

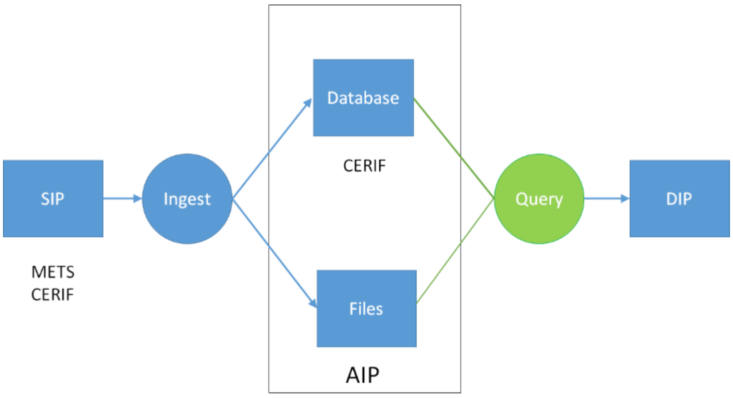
UUID_UNI = UUID_V5(UUID_DNS, 'uni.edu')
UUID(object) = UUID_V5(UUID_UNI, uni.edu_identifier(object))

The addition of globally unique identifiers - on both entities as well as relations - serves two purposes. Firstly, if all objects have globally unique identifiers the ingest process can be streamlined and do not require the full XML to be processed as a whole hence enabling CERIF XML to be used in large scale context which can otherwise be problematic[8]. Secondly, with identifiers on all objects tracking of the provenance can be made on all entities hence enabling harvesting the same object from different sources without causing the data to be mixed and corrupted. This is especially important in the scope of rights and permissions.

Using UUID identifiers also facilitates GDPR[9] compliance as it enables the involved archives to minimize the amount of personal information which they have to store and use locally and can in many case be regarded as pseudonymization by design.

Transfers

CERIF modeled data can be easily coded as XML[10]. One or several CERIF XML files can be packed into a METS [11] structure as data, where relevant metadata for the transfer is added to the corresponding sections[12]. The advantage of having the whole archive modeled as CERIF is that both the structure of the archive and the archival packages have the same format and can be seamlessly added to each other to form bigger archives.



Figur 3. Both the archive (AIPs) as well as the ingest process (SIPs) representation of the data can use CERIF and hence greatly reduce the complexity of the Ingest and Query functionality enhancing the findability (FAIR!).

Accessibility

Archived data might have access restrictions. In order not to force the whole archive to be restricted, an entity level classification can be added to prevent export or access to specific entities. More complicated access rights descriptions can be added to a Premis structure in a cfMeasurement attached to the entity. Being able to apply accessibility constrains at an entity level allows more accessibility to the open parts as there is no need to close the whole archive just because it contains some information that should not be accessible. (FAIR!)

We give an example of this, related to our case with archiving student theses. Normally, these theses should be published online as soon as they are archived, so that anyone on the Internet can retrieve them automatically in a DIP from the archive. However, some of the theses have an embargo for publishing. They still have to be in the archive, so that they can be made available upon request, but they must not be included in a publicly accessible DIP. We show how this can be handled using access restrictions of the mentioned type.

Example

About 1000 student theses per year are produced at TheUniversity. They are public records, and thus have to be archived in accordance with Swedish law, and regulations from the Swedish National Archives[6]. Moreover, most theses are published online as soon as they are deposited.

At TheUniversity, the different departments are considered archive creators on their own, and they have archived the theses independently of each other. It is not fully satisfactory to have what is essentially one central process at the university divided this way. We show how a CERIF-based database can be used to manage an archive with the examination of students as one unified process which, at the same time, can be connected with the individual departments. We then give examples of interoperability with different transmission and metadata standards, such as METS and EAD, for transferring archival descriptions to other systems.

Summary

An archival structure based on a cfProject tree is proposed. Archived objects are represented by cfResult* entities and their descriptive metadata is given in attached Cerif entities. Additional archival metadata is stored in Premis format inside attached cfMeasurment entities.

When CERIF is employed in relevant archive processes, a FAIR compliant archive is easier to achieve.

References

1. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018.
2. Lavoie, B., *Meeting the challenges of digital preservation: The OAIS reference model.* 2000.
3. euroCris, *Main features of CERIF.*
4. euroCris, *CERIF-1.6.*
5. Kunis, R.R., G.; Schwind, M. *A new Model for Document Management in E-Government Systems Based on Hierarchical Process Folders*. in *Proc. of the 7th European Conference on e-Government*. 2007. Academic Conferences Limited.
6. Swedish National, A., *RA-FS 2008:4: Föreskrifter om ändring i Riksarkivets föreskrifter och allmänna råd (RA-FS 1991:1) om arkiv hos statliga myndigheter*. 2008.
7. Library of, C., *PREMIS*. 2017.
8. Vestdam, T., B. Plauborg, and L. Van Campe, *FRIS R3 - CERIF XML in Large Scale Exchange of Research Information.* Procedia Computer Science, 2017. **106**: p. 74-81.
9. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* Official Journal of the European Union,. **L119** (4 May 2016): p. 1-88.
10. *CERIF 1.5 XML Data Exchange Format Specification*. [cited 2018 2018-03-16].
11. Library of, C., *Metadata Encoding & Transmission Standard.* 2017.
12. Gartner, R., *The Digital Object in Context: Using CERIF With METS.* Journal of Library Metadata, 2012. **12**(1): p. 39-51.