Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 146 (2019) 190–200

Procedia Computer Science

www.elsevier.com/locate/procedia

14th International Conference on Current Research Information Systems, CRIS2018

A FAIR archive based on the CERIF model

Alejandro Engelman[a]*, Christer Enkvist[a], Karl Pettersson[b]

[a]*University Library, Swedish University of Agricultural Sciences, Box 7071, Uppsala, SE-75007, Sweden*
[b]*Vice-Chancellor's Office/Documentation and Legal Affair,Swedish University of Agricultural Sciences, Box 7070, Uppsala, SE-75007, Sweden*

Abstract

The Common European Research Information Format (CERIF) has been developed as a flexible model for all kinds of research data. The focus has so-far been primarily on the construction of CRIS systems and inter-system data transfer. In the article we describe how the CERIF model can be used in the construction of an electronic archive. The advantage of basing the archive on CERIF is that an archive compliant to the FAIR guiding principles is easier to archive, in particular in areas concerning issues regarding findability, re-usability, and accessibility. In our example we represent the archive tree by cfProjects and the archived objects by cfResult* entities with their descriptive metadata given in attached CERIF entities. Archival preservation metadata is stored in the Premis format inside attached cfMeasurement entities. An example in which EPrints repository items are transferred to the archive is also presented.

Peer-review under responsibility of the scientific committee of the 14th International Conference on Current Research Information Systems, CRIS2018.

Keywords: CERIF; OAIS; Archival structure; Submission agreement; FAIR; GDPR

* Corresponding author. Tel.: +46705863108
E-mail address: alejandro.engelmann@slu.se
Supplemental information can be downloaded from Zenodo [1]

Peer-review under responsibility of the scientific committee of the 14th International Conference on Current Research Information Systems, CRIS2018.
10.1016/j.procs.2019.01.076

1. Introduction and background

Electronic archives are, in many cases, created in systems that are aimed for specific activities without interoperability in mind. The standard model for digital preservation systems today is the Open Archival Information System reference model (OAIS) [2]. OAIS is based around three information stages, Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP). SIP is the information package which is submitted to the archive, AIP the information within the archive, and DIP the information delivered from the system. The AIP is naturally dependent on the recipient's access rights to the data and parts thereof. The metadata focus of the OAIS model as well as and most real-life archives concerns the archive packages and not the explicit content of these. We find this problematic in the scope of the FAIR [3] principles as from a user/recipient perspective the findability, accessibility, and re-usability concerns the data within the AIP rather than the AIP itself.

The Common European Research Information Format (CERIF) [5, 6] has been developed as a flexible model to describe (any) research data and information, both as a database model but also as a transfer method between repositories. The CERIF model is based around base entities and relations where both the base entities as well as the relations can have associated attributes and roles. For example, a Person (cfPerson entity) can be author (role) of (relation) a publication (cfResPubl entity). This flexibility provides a dynamic framework implementable on many scopes but remaining easy to parse for information without explicit beforehand information on the contents of the data package or database model.

Archives can be considered as being a repository with some special requirements and constraints, for example the need to ensure long time usability of the deposited data. This article discusses how a CERIF based archive structure will manifest itself in the different information package stages of the OAIS model and advocates that if CERIF is employed in relevant archive processes, a FAIR [3] compliant archive can be easier to achieve. Finally, we present a real-life example from the Swedish University of Agricultural Sciences (SLU) where CERIF is employed in the archive process of student theses.

2. Metadata standard

In the framework of archives there are two fundamental scopes regarding metadata, information regarding the data object and metadata on the contents of the object. That is, we should be able to describe not only the data object at hand, for example, a project with name, team members and funders, but also the associated data such as resulting publications and data sets. Furthermore, the choice of metadata standard should be flexible enough to provide the sub archives (data providers) to extend their scope without the overly complicated discussion with the central archive. An example of this is that a sub-unit of a university is adding new archives or begins using new equipment when creating data. All these changes should be easy for the main archive to incorporate in the AIP as well as in the actual database structure of the archive so that querying and producing DIP for end users is possible in a consistent way.

In the initial FAIR guiding principles, the focus has been on data and metadata. However, in order to apply FAIR principles on more complex collections, as those from Libraries, Archives, and Museums, three main levels of information should be considered [4]:

1. Objects (such as books, journals, artefacts, videos, datasets, etc.)
2. Metadata about the objects on elementary level (such as title, creator, identifier, date, etc.)
3. Metadata records (body of metadata elements about an object in a specific database)

The levels above can also be as a nesting of information, data is both data as well as metadata. Furthermore, the FAIR principles should be invoked on all three levels which means that regardless of layer {Object, Metadata, Metadata record} we should at least have

- Globally persistent identifiers (Findable)
- Available metadata (Findable)
- Access, use and license information (Re-useable)

The most common way to accommodate several layers of information is to have different metadata standards for each level. A problem with specific metadata structures depending on the metadata records is that the parsing of an information package to large extent often requires pre-hand information of the data, or at least metadata structure and standard. CERIF, in contrast with other metadata standards with more localized focus, provides a dynamic and expandable framework to describe the whole research process which is paramount for many archives. This means that CERIF should be an excellent candidate to describe metadata in all three layers of information for many collections, and archives in particular, in order to simplify FAIR compliance.

3. CERIF as an archival model

In order to build an archive structure based on CERIF we use properly classed cfProject entities to construct the archive tree. These entities represent real processes, projects or activities that produce content or groupings of them in accordance with current archival recommendations that documents should be classified from a process perspective [7, 8]. Naturally, the cfProject entity can also be an 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, like contributors (cfPerson), organisations (cfOrganisationUnit), fundings (cfFunding) and other properly classed CERIF entities. The files would be represented by cfMedium objects and the file itself would be included in the submissions and ingested to the archive.

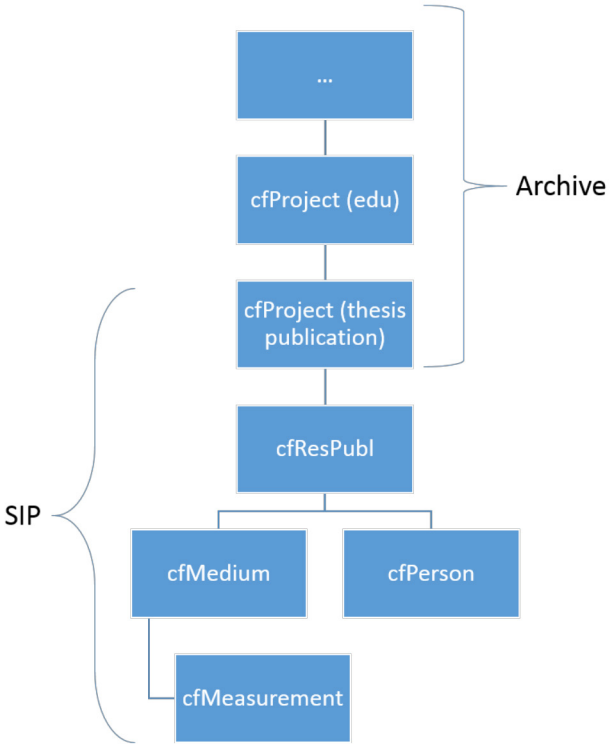

Fig. 1. The process archive of the student theses in CERIF terms. The main process is "education processes" (edu), sub-process "thesis publication". The published result is typically a PDF-file (cfMedium) with associated cfMeasurement containing embargo (Premis) information.

Master data like the organisation structure, lists of persons, 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 [9] XML data is stored. The object identifiers in the Premis model are equated to the CERIF entity identifiers, making the reference to them transparent.

```
<rights>
  <rightsStatement>
    <rightsStatementIdentifier>
      <rightsStatementIdentifierType>UUID</rightsStatementIdentifierType>
      <rightsStatementIdentifierValue>f66ca00a-0759-5465-a3c7-f07c3cdef9be</rightsStatementIdentifierValue>
    </rightsStatementIdentifier>
    <rightsBasis>Policy</rightsBasis>
    <rightsGranted>
      <act>Disseminate</act>
      <restriction>Allow</restriction>
      <termOfGrant>
        <startDate>2019-01</startDate>
        <endDate>open</endDate>
      </termOfGrant>
    </rightsGranted>
    <linkingObjectIdentifier>
      <linkingObjectIdentifierType>cfMediumId</linkingObjectIdentifierType>
      <linkingObjectIdentifierValue>fec6df47-5c30-5ab4-a54d-2af558c84f24</linkingObjectIdentifierValue>
    </linkingObjectIdentifier>
  </rightsStatement>
</rights>
<!--Embargo date: 2019-01-->
</premis>
```

Fig. 2. Example of a rights section in a PREMIS representation describing an embargo restriction (the complete file can be downloaded from Zenodo [1].). The PREMIS information can either be a separate file in the SIP or contained in the CERIF XML file in a cfMeas block.

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!).

Using CERIF in the whole archive process have further advantages as the access information, on a very granular level, can be retrieved directly from the CERIF XML file, hence easily implemented in the archive database and facilitating queries where the result is a correct DIP (taking the queries' permissions into consideration).

4. 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.

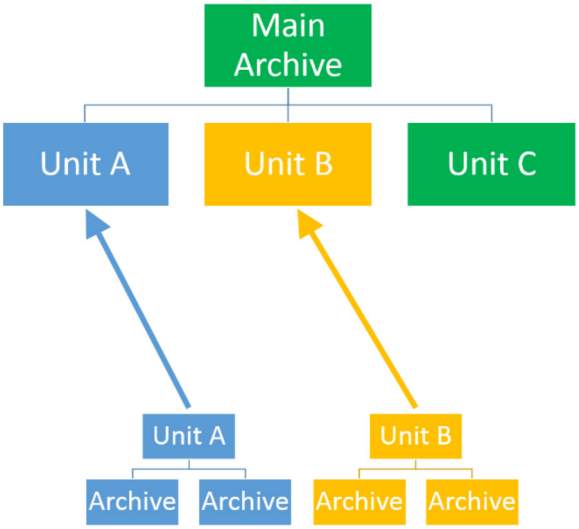

Fig. 3. The archive process should be close to the organisational unit producing the actual content. If the involved units transfers their archive information in CERIF the main archive will have a trivial task of organising the information and extending the archive tree.

In the submission agreement the main archive simply have to provide the identifier of the sub-units top-level cfProject, the sub-unit can thereafter simply provide the structure of the sub-archive in the CERIF XML file.

5. Identifiers

Every CERIF entity has an identifier. Although it is not mandatory with globally unique identifiers, UUIDs are often used. Globally unique identifiers, or at least persistent identifiers, is also important for the data to be FAIR [10]. 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 the Swedish University of Agricultural Sciences (SLU) is to use version 5 UUID identifiers constructed from, where applicable, local identifiers combined with a UUID of the domain in which they are unique. For a SLU-identifiers the domain name slu.se is normally as initial name space.

```
UUID_SLU = UUID_V5(UUID_DNS, 'slu.se')
UUID(object) = UUID_V5(UUID_SLU, slu.se_identifier(object))
```

When an object has no institution-wide unique identifier but only a local identifier, a local UUID namespace is constructed, for example with the identity provider URL and that is used to produce globally unique identifiers.

```
UUID_local = UUID_V5(UUID_DNS, 'local.slu.se')
UUID(object) = UUID_V5(UUID_local, local_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 [11]. 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 [12] compliance as it enables the involved archives to minimize the amount of personal information which each archive unit have to store and use locally and can in many cases be regarded as pseudonymization/privacy by design. For sensitive information the identifiers should naturally be replaced with encrypted identifiers rather than the hashed UUIDs discussed above.

6. Transfers

CERIF modelled data can be easily coded as XML [13]. One or several CERIF XML files can be packed into a METS [14] structure as data, where relevant metadata for the transfer is added to the corresponding sections. One advantage of having the whole archive modelled 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. The addition of a new submission to the archive is simply an addition of a cfProject* tree.

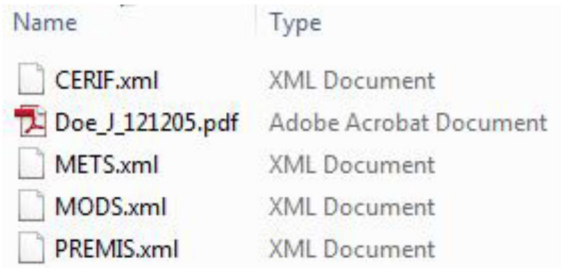

Fig. 4. Example of a SIP consisting of the thesis as PDF and with metadata in CERIF XML, METS XML, MODS XML, and Premis XML files. A complete example of the included XML files can be downloaded from Zenodo [1].

There are many transfer protocols with similar constructions, one example is the common specifications for publications (FGS-PUBL) [16, 17] describing publication submissions to the Swedish National Library. A CERIF based SIP is exemplified in Fig. 4 where a MODS XML file has been added to make the SIP compatible with the FGS-PUBL. Furthermore, in the case that CERIF is inappropriate model for the metadata at hand the MODS file may be used as information container.

7. Accessibility and Findability

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!). The complete archive process is schematically drawn in Fig. 5.

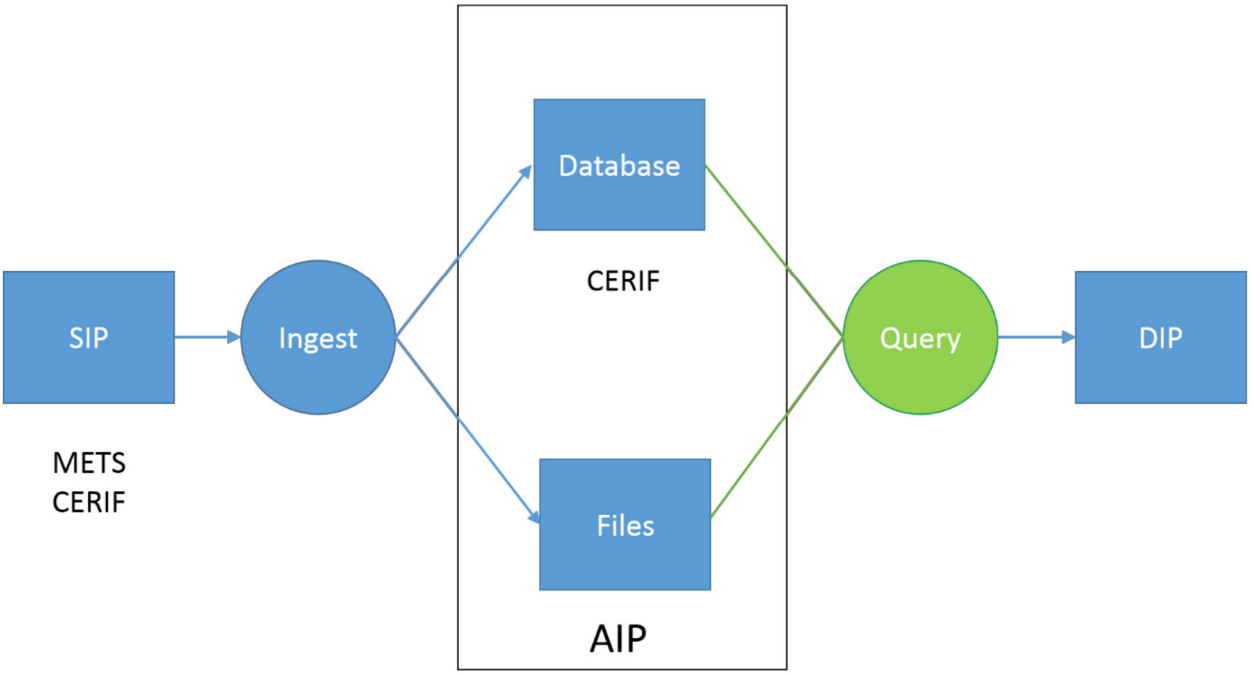

Fig. 5. Both the archive (AIPs) as well as the ingest process (SIPs) representation of the data can use CERIF and hence greatly reduce the complexity of the Ingest and Query functionality enhancing the accessibility as well as findability (FAIR!).

The search of the database is limited by the amount and form of metadata stored during the ingest procedure. As the SIPs can contain a variety of metadata schemas it is difficult to arrange a database structure to accommodate them. We propose that as much metadata as possible is included in the SIP in CERIF format. In the case of research output the CERIF schema is expected to cover a substantial part of the metadata. Other metadata schemas can be delivered together with the CERIF information if additional metadata is available. The CERIF metadata which as required by the submission agreement contains only globally unique identifiers can be readily be added to the CERIF database under the Data Management function of the OAIS schema.

An example of accessibility restrictions, related to our case with archiving student theses is that these theses should normally 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.

8. Example

About 1000 student theses per year are produced at the Swedish University of Agricultural Sciences (SLU). They are public records, and thus has to be archived in accordance with Swedish law, and regulations from the Swedish National Archives [8]. Moreover, most theses are published online in an open access repository located at SLU as soon as they are deposited [18].

At SLU, 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.

The SLU open access repository for student theses uses EPrint [19] but other repositories as for example DiVA [20, 21] employs similar setups where a SIP is harvested from the repository. Naturally, having a CERIF based repository simplifies the construction of the CERIF XML used in the SIP but this is naturally not mandatory.

Fig. 6. Example of a METS file. Since almost all metadata can be located in the CERIF file the METS file can be quite compact. The complete file can be downloaded from Zenodo [1].

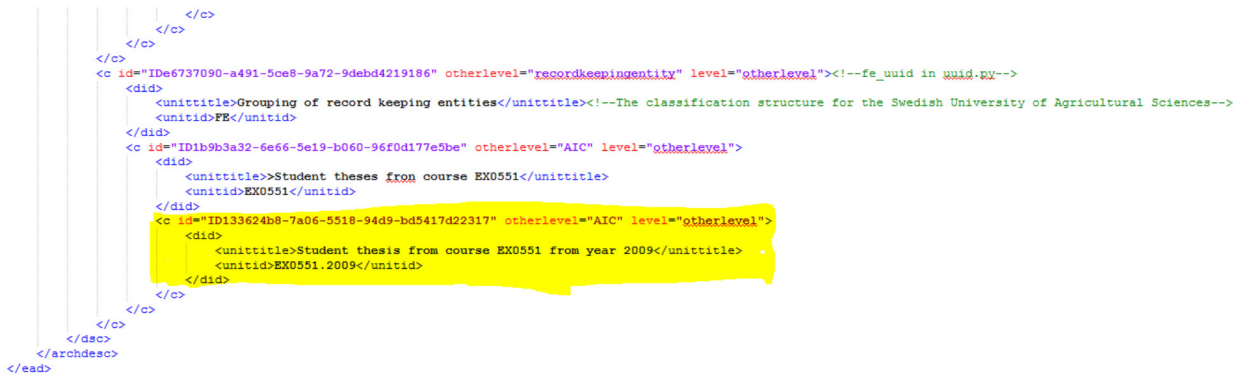

Fig. 7. In the EAD file the archive is provided with the identifiers to represent the top level cfProject, in the highlighted example student theses from a specific course and year. The complete file can be downloaded from Zenodo [1].

Each department is in the submission agreement provided with an identifier for their cfProject entry in the main archive. An entry in the EAD XML file for the sub-unit containing theses from 2009 is shown in Fig. . Although not commonly used, nothing prohibits that the department itself adds sufficient information regarding their sub-series in

the SIP delivery so that the main archive can extend the classification structure in the EAD as the departments creates new units. Empowering the sub-archives to provide the archive structure greatly facilitates the main archives burden to keep the EAD updated, especially in the cases where sub-archives are created on a regular basis, in the SLU example is a sub-archive created for the combination of course and academic year.

9. 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. By representing as much metadata as possible in CERIF and storing it in a searchable CERIF database, findability increases as queries can be placed on the archive independently of any specific organization chosen for the data management besides the CERIF backbone. Being able to invoke fine-grained access rights the demise "as open as possible, as closed as necessary" can be followed instead of placing restrictions on the whole of the archive fulfilling the FAIR accessible criteria. Interoperability is supported as in the same way as the archive is built by CERIF organized deliveries (SIP), the information can be sent to other recipients in the same orderly way (DIP). Finally, with extensive use of globally unique identifiers a CERIF entity might be placed in several different contexts retaining its identity allowing the transfer and reuse in a controlled way.

References

1 Engelmann, A., Enkvist, C. & Pettersson, K. *Example files to A FAIR archive based on the CERIF model*, <https://dx.doi.org/10.5281/zenodo.1287781> (2018).

2 Lavoie, B. Meeting the challenges of digital preservation: The OAIS reference model. *OCLC Newsletter* **243**, 26-30 (2000).

3 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, doi:10.1038/sdata.2016.18 (2016).

4 Lukas Koster, Saskia & Woutersen-Windhouwer. FAIR Principles for Library, Archive and Museum Collections: A proposal for standards for reusable collections, *The code4lib journal*. **40**, < http://journal.code4lib.org/articles/13427> (2018).

5 euroCris. *Main features of CERIF*, <https://www.eurocris.org/cerif/main-features-cerif> (

6 euroCris. *CERIF-1.6*, <https://www.eurocris.org/cerif/feature-tour/cerif-16> (

7 Kunis, R., Rünger, G. & Schwind, M. in *Proc. of the 7th European Conference on e-Government.* 229-240 (Academic Conferences Limited).

8 Swedish National Archives. *RA-FS 2008:4: Föreskrifter om ändring i Riksarkivets föreskrifter och allmänna råd (RA-FS 1991:1) om arkiv hos statliga myndigheter*, <https://riksarkivet.se/rafs?item=104> (2008).

9 Library of Congress. *PREMIS*, <http://www.loc.gov/standards/premis/> (2017).

10 Hodson, S. *et al.* FAIR Data Action Plan. Interim recommendations and actions from the European Commission Expert Group on FAIR data. doi:10.5281/zenodo.1285290.

11 Vestdam, T., Plauborg, B. & Van Campe, L. FRIS R3 - CERIF XML in Large Scale Exchange of Research Information. *Procedia Computer Science* **106**, 74-81, doi:10.1016/j.procs.2017.03.037 (2017).

12 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** 1-88.

13 Dvořák, J. & Jörg, B. *CERIF 1.5 XML Data Exchange Format Specification*, <http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF1.5_XML.pdf> (2013).

14 Library of Congress. *Metadata Encoding & Transmission Standard*, <http://www.loc.gov/standards/mets/> (2017).

15 Gartner, R. The Digital Object in Context: Using CERIF With METS. *Journal of Library Metadata* **12**, 39-51, doi:10.1080/19386389.2012.661689 (2012).

16 Swedish National Archives. *Förvaltningsgemensamma specifikationer (FGS)*, <https://riksarkivet.se/fgs-

earkiv> (2018).
17 Swedish National Library. *Common Specification (FGS-PUBL) for deposit of single electronic publications to the National Library of Sweden - Kungl. biblioteket (KB)*, <http://www.kb.se/namespace/digark/deliveryspecification/deposit/fgs-publ/v1/FGS-PUBL_1_1_eng.pdf> (2016).
18 *Epsilon Archive for Student Projects*, <https://stud.epsilon.slu.se/> (
19 Brody, T. *et al. EPrints 3.3.16*, <http://files.eprints.org/2306/> (2017).
20 DiVA. *Om DiVA*, <https://wiki.epc.ub.uu.se/display/divainfo/Om+DiVA> (
21 DiVA. *DiVA Archive Package Delivery Service*, <https://wiki.epc.ub.uu.se/display/divainfo/DiVA+Archive+Package+Delivery+Service> (2018).