Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

OpenAIRE Guidelines for CRIS Managers: Supporting Interoperability of Open Research Information through established standards

Nikos Houssos[a], Brigitte Jörg[b], Jan Dvořák[c], Pedro Príncipe[d], Eloy Rodrigues[d], Paolo Manghi[e], Mikael K. Elbæk[f]

[a]*National Documentation Centre / National Hellenic Research Foundation, Greece, nhoussos@ekt.gr*
[b]*JeiBee Ltd., United Kingdom, brigitte.joerg@gmail.com*
[c]*Institute of Information Studies and Librarianship, Faculty of Arts, Charles University in Prague, Czech Republic, jan.dvorak@ff.cuni.cz*
[d]*University of Minho, Portugal, {pedroprincipe, eloyrodrigues}@sdum.uminho.pt*
[e]*Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", paolo.manghi@isti.cnr.it*
[f]*Technical Information Center of Denmark, Technical University of Denmark, Denmark, miel@dtic.dtu.dk*

Abstract

OpenAIRE is the European infrastructure enabling researchers to comply with the European Union requirements for Open Access to research results. OpenAIRE collects metadata from data sources across Europe and beyond and defines interoperability guidelines to assist providers in exposing their information in a way that is compatible with OpenAIRE. This contribution focuses on a specific type of data source, CRIS systems, and the respective OpenAIRE guidelines, based on CERIF XML. A range of issues, spanning different aspects of information representation and exchange, needed to be addressed by the guidelines in order to define a complete solution for interoperability.

Peer-review under responsibility of euroCRIS.

Keywords: interoperability; research information; open access; OpenAIRE; CRIS; CERIF; OAI-PMH.

1. Introduction

OpenAIRE supports the European Commission Open Access policy by providing an infrastructure for researchers to comply with the corresponding requirements for Open Access to research results[1,2]. The 2nd generation of the OpenAIRE infrastructure is currently being developed in the OpenAIREplus project to move from an infrastructure focusing on publications by EU funded projects to a more comprehensive infrastructure that covers

Peer-review under responsibility of euroCRIS.

all types of scientific output supported by any type of funding. More specifically, it will facilitate access to the entire Open Access scientific production of the European Research Area, providing cross-links from publications to datasets and funding schemes, including national programmes. To this aim the infrastructure provides services to (*i*) continuously harvest from data sources (e.g. literature repositories, data repositories, CRIS systems) metadata descriptions relative to research outcome as well as the full-text of publications (the latter currently used for the automatic mining of links between publications and datasets, projects and research programmes and initiatives), and (*ii*) aggregate, harmonize and enrich by inference such content.

In such a scenario, one of the missions of OpenAIRE is to facilitate and encourage data exchange between European research-oriented information providers. To put this into practice, an integrated suite of *guidelines for data sources* was developed, with the purpose of recommending a common and interoperable approach for exporting metadata among each class of data sources. As a result, this will facilitate data exchange and significantly increase the exposure and visibility of a wide range of information providers in Europe and beyond. The OpenAIRE infrastructure defined and currently disseminates guidelines addressing the three classes of research-oriented data sources relevant to its mission: Literature repositories, Data archives and CRIS systems. The guidelines target managers of these types of systems and generally specify: (*i*) standard protocol and relative configuration to be used for exporting metadata, and (*ii*) the XML data format (hence the corresponding data model) to be used to export metadata. Dissemination and taking-up of the guidelines will ensure compatibility to the OpenAIRE infrastructure, i.e. ease the integration of content in its aggregation and associated increase of visibility for data sources, but will also promote interoperability between such data sources and other research infrastructures, for the benefit of the knowledge community at large.

In an effort to make the OpenAIRE guidelines as open as possible to the wider public, OpenAIREplus have established a specific Wiki[3] with all the needed information about the three aforementioned guidelines. The goal of this wiki is to provide a public online space to share OpenAIRE's work on interoperability and to engage with the community, posting experiences, adding comments, and promoting for all the OpenAIRE data providers.

The aim of this paper is to elaborate on the principles adopted in the definition of the *OpenAIRE guidelines for CRIS managers* to better understand and appreciate the choices behind the solution provided. The structure of the rest of the article is the following: Section 2 provides an overview and scope of the OpenAIRE guidelines for CRIS managers. Section 3 describes in detail how the guidelines address the different aspects of interoperability, leading to a complete solution for CRIS managers. The article concludes with a summary – future work section.

2. OpenAIRE Guidelines for CRIS Managers: aims and scope

The OpenAIRE infrastructure is today populating a graph of entities that includes publications, datasets, projects, organizations, data sources, and persons. Its data model is strongly inspired by the CERIF model[4,6]. In this sense it not only includes a subset of CERIF entities, but also the notion of "semantic layer"[5]. As such, the OpenAIRE infrastructure can be itself considered a CRIS system. Such a choice facilitated the process of definition of the guidelines for CRIS managers, which adopt CERIF XML as the basis for harvesting and importing metadata from CRIS systems and essentially incarnate a specialisation of CERIF for OpenAIRE. The guidelines provide orientation for CRIS managers to expose their metadata in a way that is compatible with the OpenAIRE infrastructure, i.e. easy to collect and easy to interpret. By implementing them, CRIS managers support smooth inclusion and therefore the reuse of metadata in their systems within the OpenAIRE infrastructure. The guidelines play therefore a twofold role: on the one hand they provide guidance for developers of CRIS platforms on how to make their CRIS system compatible with OpenAIRE; on the other hand they describe how third-party consumers can access and interpret content from an OpenAIRE-compatible CRIS system. Worth of notice is that the information exchange among individual CRIS systems and OpenAIRE is an example of point-to-point data exchange between CRIS systems.

CERIF is a data model for research information. It allows for the formal description of many aspects inherent in the research domain, covering the entire research lifecycle. Its aim is to enable interoperability and data exchange between systems managing research information. The OpenAIRE infrastructure is one of such systems, as its information space data model can be described as a subset of the CERIF data model.

The OpenAIRE Guidelines for CRIS managers provide a holistic solution for the exchange of research information among individual CRIS systems and OpenAIRE. Different levels of interoperability needed to be considered and addressed, namely *system*, *syntax*, *structure* and *semantics*[7,8].

Therefore, the definition of the OpenAIRE Guidelines for CRIS managers can be considered to comprise work along two dimensions:

- *Structure* and *semantics*. This dimension is concerned with achieving structural and semantic interoperability. It includes the identification of the CERIF data model subset that is relevant to OpenAIRE and the vocabularies of the CERIF semantic layer that apply to OpenAIRE. The result is a conceptual specialisation of the CERIF model with specific semantics for the purposes of the OpenAIRE guidelines and could be termed the *OpenAIRE-CERIF Profile*, since this type of artefact is sometimes called a *Profile* in the CRIS community; however there is no broad consensus yet in the CRIS community on the exact meaning and suitability of this terminology.
- *System* and *syntax*. This dimension concerns the system and syntax interoperability and addresses issues of access to data encoded in CERIF XML through OAI-PMH.

These dimensions of the guidelines definition and the design principles and choices employed for them are elaborated in the following section.

3. OpenAIRE Guidelines for CRIS Managers: addressing difference aspects of interoperability

The present section describes how the different aspects of interoperability (structure and semantics, system and syntax) are addressed in the OpenAIRE guidelines for CRIS Managers.

3.1. The definition of the OpenAIRE-CERIF data model structure and semantics

Fundamental prerequisites for achieving interoperability are the definition of the structure and semantics of the data to be exchanged between systems. The current section elaborates on these two issues.

- *Structural interoperability* concerns the conceptual and logical representation of the information to be transferred among different systems, for example, which data elements are being used to represent information and what constraints apply. In the case of the OpenAIRE CRIS guidelines the major part of this work has been performed during the initial definition of the enhanced OpenAIRE data model within the OpenAIREplus project, where OpenAIRE entities and attributes have been mapped to a subset of CERIF and the CERIF semantic layer has been used to represent relationships among entity instances. The list of simple CERIF entities (excluding multi-lingual field entities, link entities and federated identifiers) relevant to OpenAIRE is the following: cfProject; cfPerson; cfOrganisationUnit; cfResultPublication; cfResultProduct; cfFunding; cfService; cfPhysicalAddress; cfElectronicAddress – furthermore federated identifiers are relevant for OpenAIRE. Integrity constraints apply inherently due to the adoption of a relational, CERIF-compliant data model and any specialisation of constraints is such that the constraints of the canonical CERIF model hold.
- *Semantic interoperability* concerns the meaning of the exchanged information. Regarding the scope of entities used in the OpenAIRE data model, the CERIF definitions for each entity in the CERIF standards semantics apply, albeit with certain specialisations that are explicitly mentioned in the guidelines (e.g. for example only a narrow selection of services represented by the cfService entity in individual CRIS systems are relevant for OpenAIRE). A very important aspect of semantics are the controlled vocabularies and terms (classification schemes and classifications in CERIF terminology) that are used in (a) classifications of entity instances (e.g. publications types, data set types) and (b) relationships among entity instances (e.g. the possible roles for an organization in a project). In technical terms, a specification of a CERIF Semantic Layer instance (i.e. a set of vocabularies) has been defined for the particular needs of OpenAIRE. The use of the specified vocabularies is mandatory, i.e. no other vocabularies can be used in the CERIF XML information exposed by CRIS systems to the OpenAIRE infrastructure. The adopted vocabularies depend on design choices of the OpenAIRE data model and are not restricted by the standard CERIF Semantics specification. However, where possible, elements of the

standard CERIF Semantics specification have been used. This demonstrates an important feature of CERIF, the extensibility in terms of semantics: CERIF and in particular the CERIF semantic layer allows the introduction of domain or application specific semantics in the data model while maintaining intact the structure of the model and hence requiring no changes in the implementation at the physical / system level.

3.2. The definition of the OpenAIRE-CERIF XML and the corresponding access protocol

Below the structure and semantics level of abstraction lie the syntax of the exchanged data and the protocol used to access it. The approach followed for these issues in the OpenAIRE CRIS guidelines is presented in this section.

- *Syntax/format*: CERIF XML has been adopted by OpenAIRE as the basis for harvesting and importing metadata from CRIS systems. It was necessary to identify the CERIF XML subset suitable for representing information for harvesting by OpenAIRE. Therefore, a separate XML namespace with its own XML Schema (the OpenAIRE CERIF XML Schema) has been defined[9].
- *Access protocol*: OAI-PMH has been selected as the access protocol for harvesting information from CRIS systems by OpenAIRE, due to its ubiquity and ease of implementation for individual systems. A specification of how OpenAIRE-CERIF XML data, essentially a graph-based structure, are to be exported via OAI-PMH protocol as sets of XML records. A detailed approach (including examples) is prescribed in the guidelines in order to specify how the fully connected graph structure underlying any CERIF/CRIS system can be marshalled in OAI-PMH compliant lists of metadata records.

In the rest of this section, we provide additional details behind the definition of the OpenAIRE-CERIF XML and the OAI-PMH configuration necessary to expose CRIS system content.

3.2.1. Format and syntax: OpenAIRE CERIF-XML

The OpenAIRE XML Schema is designed so that any information encoded in OpenAIRE-CERIF XML also conforms to the standard CERIF XML schema (modulo namespace). The opposite does not hold, i.e. there can be valid standard CERIF XML that does not conform to the OpenAIRE CERIF XML schema. This was achieved by adhering to the following two principles:

- All data elements that can be included in the CERIF XML data for OpenAIRE are part of the standard CERIF XML specification. There is no entity, attribute or relationship in the OpenAIRE CRIS guidelines that does not exist in standard CERIF XML.
- All constraints (e.g. cardinality) enforced through the standard CERIF XML Schema are also valid in the OpenAIRE CERIF XML.

The CERIF XML style allowed is the one defined in CERIF 1.6 XML specification, where the following general rules hold (reference to the full specification[10,11] is recommended for details):

1. The CERIF data must be represented as descendants of a root XML element, "CERIF".
2. Direct descendants of the CERIF elements must be only simple CERIF Entities (such as Person; Project; OrganisationUnit; ResultPublication; ResultProduct) not multi-lingual or link entities or federated identifiers.
3. The OpenAIRE-CERIF XML of any simple CERIF entity (*i*) embeds multilingual attributes, link entities and federated identifiers, (*ii*) cannot be embedded another simple CERIF entity, and (*iii*) may embed links to other entities (only links, not the entire record of the entity on the other end of the relationship).
4. Every classification and classification scheme used in link entities should belong to the set of classifications and classification schemes (i.e. vocabularies) that constitute the OpenAIRE-CERIF profile specification.
5. Referential integrity constraints for all relationships among entities should be satisfied in the OpenAIRE-CERIF XML records exported by the CRIS system. In particular, it is required that if two CERIF objects bear a relationship between them, the OpenAIRE-CERIF XML records of the two objects (even if linked records are retrieved as part of different OAI-PMH sets, see Section 3.3) **must** bear bi-lateral relationships. For example,

consider the case of a relationship between *cfOrgUnit A* and *cfProject B*. The hosting CRIS system must export two OpenAIRE-CERIF XML records:

a. An XML record for cfOrgUnit A. This XML record must contain, as a nested XML element, the cfProj_OrgUnit linking entity instance connecting cfProj B.
b. An XML record for cfProj B. This XML record must contain, as a nested XML element, the cfProj_OrgUnit linking entity instance connecting cfOrgUnit A.

It is worth noting that the two aforementioned XML records (cfOrgUnit A and cfProj B) may be contained in distinct sets of XML records exported by the CRIS system through separate OAI-PMH sets (see Section 3.2.2).

The referential integrity rule is mandatory. Achieving referential integrity is greatly facilitated by the inherent integrity of the CERIF data model and the adoption of a CRIS system platform that uses a native CERIF database.

3.2.2. Transmission of CERIF XML as OAI-PMH payload

OAI-PMH is a widely adopted standard protocol for exporting metadata records from a data source. An OAI-PMH exposes a set of records; each record must be in XML format and use UTF-8 encoding. The protocol delivers (i) paging mechanisms ("resumption token"), to enable controlled, incremental access over arbitrary large collections of metadata records, and (ii) selective harvesting based on timestamps or set-based access ("OAI Sets"), to enable partitioning of the information space of records into meaningful subsets (data source specific). Originally conceived to support bibliographic metadata exports, due to its simplicity OAI-PMH has been subsequently adopted also in related contexts. For the same reasons, the OpenAIRE CRIS guidelines for CRIS managers adopted OAI-PMH as the means to deliver records to the OpenAIRE infrastructure.

CRIS systems implement the CERIF data model. As such, they store in their back-ends a graph of interconnected objects. In particular, objects belong to a given CERIF entity and may bear relationships to other objects of the same entity or different entities. Therefore, a way to serialise the graph of CERIF XML records is required to enable transmission via OAI-PMH. The approach specified by the guidelines states that the OAI-PMH protocol must expose OpenAIRE-CERIF XML records according to the following configuration:

- Each simple entity in the OpenAIRE-CERIF profile has an associated OAI-PMH set: cfProject; cfPerson; cfOrgUnit; cfResultPublication; cfResultProduct; cfFunding; cfService;
- Objects following a simple CERIF entity are exported via the relative OAI-PMH set as records conforming to the relative OpenAIRE-CERIF XML;
- One special OAI-PMH set must be provided, from which the XML records of all objects, independently from their entity, can be collected. This set essentially provides the whole serialized graph of the CERIF XML information as a single set of XML records.

Therefore, a typical sequence of steps for a harvester to retrieve the entire set of records in a source CRIS system is the following:

1. Retrieve all records in each of the seven sets corresponding to the seven simple CERIF entities.
2. Before transferring the information to the target system perform integrity checks on all the data records.
3. Transfer the checked records to the target system, except problematic ones. Report any problems to the source CRIS system.

The harvester can also retrieve the whole serialized graph of CERIF XML information as a single set. However, in such cases the amount of data to be retrieved will be very large and the load on the CRIS system server to generate the whole CERIF XML of the source CRIS database might be quite high (depending on implementation). Therefore, the approach of gradually retrieving all information through entity-specific sets is foreseen to be used more frequently in practice. This is important for an initial transfer of metadata contents from a CRIS to the OpenAIRE infrastructure.

Another option provided by the OAI-PMH protocol is the selective harvesting of records based on the last modified date of each record, which enables for example the harvester to retrieve from a particular source only the

records that were updated since the last harvest. This is particularly useful for very large sources. Due to considerations regarding not entirely consistent and reliable mechanisms for setting date stamp values in certain source systems, OpenAIRE generally tends to avoid employing selective harvesting based on last update dates. If reliable mechanisms for setting date stamps are present in a source CRIS system, OpenAIRE may employ selective harvesting, for example in the case of very large data sources.

For CRIS systems that are able to reliably set last modified data values in records, the following rules apply regarding setting values of date stamps for CERIF XML records exposed by CRIS systems to OpenAIRE via OAI-PMH: Date stamps should be set by CRIS systems in records, based on the following last update principle: the date stamp should reflect the last date/time where any information contained within the record payload (e.g. entity fields, multilingual fields, federated identifiers, linked entities). Any such modification should result in a modification of the date stamp; under no circumstances can the date stamp be earlier than this date. Modifications of other entities, not included as nested content in the record payload (e.g. entity instances connected via link entities with the entity instance in question) must not result in an update of the record date stamp value.

Summary

The OpenAIRE guidelines for CRIS systems comprise a set of specifications that provide a complete solution for interoperability between individual CRIS systems in Europe and beyond and the OpenAIRE infrastructure. Essentially, this is an example of point-to-point data exchange between CRIS systems. Several different interoperability aspects needed to be addressed to achieve the purpose of the guidelines, namely access protocol, syntax, structure and semantics. The approach taken was to adopt established standards, protocols and mechanisms, such as the CERIF model and semantic vocabularies, the CERIF XML exchange format and the OAI-PMH protocol. Furthermore, to create a coherent set of specifications that are fit for purpose, certain specialisations, custom mechanisms and conventions based on existing solutions have been necessary.

Acknowledgements

The work presented in this paper has been partly supported by the OpenAIREplus Project (Ref No: 283595) of the European Union FP7-INFRASTRUCTURES Programme. Jan Dvořák's work on this article was partly supported by the Ministry of Education, Youth and Sports of the Czech Republic through grant no. LG14007. The authors wish to acknowledge the valuable feedback provided by the reviewers of the OpenAIRE guidelines for CRIS managers (the reviewers' list is available at the OpenAIRE guidelines wiki[3]).

References

1. Manghi P, Bolikowski L, Manola N, Schirrwagen J, Smith T. OpenAIREplus: the European Scholarly Communication Data Infrastructure. *D-Lib Magazine* 2012;**18**(9), 1.
2. Rettberg N, Schmidt B. OpenAIRE—Building a Collaborative Open Access Infrastructure for European Researchers. *Liber Quarterly: The Journal of European Research Libraries* 2012;**22**(3).
3. OpenAIRE Guidelines Wiki, http://guidelines.openaire.eu.
4. Jeffery K, Houssos N, Jörg B, Asserson A. Research information management: the CERIF approach. *International Journal of Metadata, Semantics and Ontologies* 2014:**9**(1):5-14.
5. Jörg B, Jeffery K, Grootel G. Towards a Sharable Research Vocabulary (SRV) – A Model-Driven Approach –. In: García-Barriocanal E, Cebeci Z, Okur M, Öztürk A, editors. *Metadata and Semantic Research*. vol. 240 of Communications in Computer and Information Science. Springer Berlin Heidelberg; 2011. p. 256-268.
6. Manghi P, Houssos N, Mikulicic M, Jörg B. The Data Model of the OpenAIRE Scientific Communication e-Infrastructure. In: Dodero JM, Palomo-Duarte M, Karampiperis P, editors. *Metadata and Semantics Research*. Springer; 2012 p. 168–80.
7. Sheth AP. Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In: *Interoperating geographic information systems*. Springer; 1999. p. 5-29.
8. Ouksel AM, Sheth A. Semantic Interoperability in Global Information Systems. *SIGMOD Rec*. 1999;**28**(1):5-12.
9. OpenAIRE CERIF XML Schema, urn:xmlns:org:eurocris:cerif-1.6-2::eu:openaire:cris-mgr-guidelines-1.0.
10. CERIF XML Data Exchange Format Specification, http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF1.5_XML.pdf.
11. CERIF XML 1.6 XML Schema, http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.6/CERIF_1.6_2.xsd.