OpenAIRE Guidelines for CRIS Managers

Version 1.0
June 2015

Abstract
The Guidelines specify the interoperability layer between Current Research Information Systems (CRIS) and the OpenAIRE infrastructure. The information interchange is based on the Common European Research Information Format (CERIF) data model, the CERIF XML exchange format, and the OAI-PMH protocol. The Guidelines are intended mainly for implementers and administrators of CRIS who plan to communicate research information to OpenAIRE. OpenAIRE (openaire.eu) is the European infrastructure enabling researchers to comply with the European Union requirements for Open Access to research results. OpenAIRE collects metadata from a variety of data sources: publication repositories, data archives and CRIS across Europe and beyond. Interoperability guidelines are defined for each type of source. CERIF is a standard data model for research information and a recommendation by the European Union to its Member States. The custody of CERIF has been entrusted by the European Union to euroCRIS (eurocris.org), an international not-for-profit organisation dedicated to the interoperability of CRIS.

Editors
Nikos Houssos (http://orcid.org/0000-0002-5277-285X)
National Documentation Centre (EKT)/NHRF, Greece
Brigitte Joerg (http://orcid.org/0000-0001-7941-8108)
JISC Innovation Support Center, UKOLN, University of Bath, United Kingdom
JeiBee Limited, London, United Kingdom
Jan Dvořák (http://orcid.org/0000-0001-8985-152X)
Charles University in Prague, Czech Republic

Experts & Reviewers
Earlier versions of these Guidelines were kindly reviewed by:
Paolo Manghi (http://orcid.org/0000-0001-7291-3210), ISTI-CNR, Italy
Mikael K. Elbæk (http://orcid.org/0000-0001-8037-7577), Technical University of Denmark, Denmark
Keith Jeffery, Keith G Jeffery Consultants, United Kingdom
Anne Asserson, University of Bergen, Norway
Andrea Bollini (http://orcid.org/0000-0002-9029-1854), CINECA, Italy
Thorsten Hoellrigl (http://orcid.org/0000-0001-6080-464X), Thomson Reuters, Germany
Teemu Kemppainen, CSC - IT Centre for Science, Finland
Thomas Vestdam (http://orcid.org/0000-0002-1336-7368), Elsevier, Denmark

Organisation names reflect the persons' affiliations at the time of their contribution to the Guidelines.

Table of Contents

1 Introduction

1.1 Aim

The Guidelines provide orientation for CRIS managers to expose their metadata in a way that is compatible with the OpenAIRE infrastructure. By implementing the Guidelines, CRIS managers support the inclusion and therefore the reuse of metadata in their systems within the OpenAIRE infrastructure. For developers of CRIS platforms, the Guidelines provide guidance to add supportive functionalities for CRIS managers and users. Exchange of information between individual CRIS systems and the OpenAIRE infrastructure is an example of point-to-point data exchange between CRIS systems, since the OpenAIRE infrastructure is itself a CRIS system.

1.2 CERIF-CRIS

CERIF (Common European Research Information Format) is a standard data model for research information and a recommendation by the European Union to Member States. The care and custody of CERIF was handed over by the European Union to euroCRIS (www.eurocris.org), a non-for-profit organisation dedicated to the interoperability of Research Information Systems (CRISs). In addition to a domain model and a formal relational (ERM) data model, CERIF defines a XML format for data exchange. The OpenAIRE data model is CERIF-compliant and CERIF XML has been adopted by OpenAIRE as the basis for harvesting and importing metadata from CRIS systems.

1.3 License

1.4 Accompanying Material

The OpenAIRE CERIF XML markup for this version of the Guidelines uses the XML namespace with the following URI:

urn:xmlns:org:eurocris:cerif-1.6-2::eu:openaire:cris-mgr-guidelines-1.0

The OpenAIRE CERIF XML Schema that defines the admissible XML structures for this XML namespace is provided at the following URL:

https://www.openaire.eu/cerif_schema/cerif-1.6-2_openaire-1.0.xsd

The OpenAIRE CERIF semantic vocabularies are attached as separate documents:

- OpenAIRE_CERIF_Semantics_v.1.0.xslx (MS Excel format)
- OpenAIRE_CERIF_Semantics_v.1.0.xml (CERIF XML format)

Examples of OpenAIRE CERIF XML embedded in OAI-PMH responses are attached in seven separate files whose names follow the openaire_cerif_xml_example_*.xml pattern. Appendix A gives an overview of the contents of the examples.

2 CERIF subset for OpenAIRE

The Common European Research Information Format (CERIF) is a comprehensive domain model that allows for the formal description of many aspects inherent in the Research domain. The current scope of the OpenAIRE information space does not cover the full range of information in CERIF, therefore it is based on a subset of the full CERIF model – see Figure 1. The Guidelines therefore focus on those information elements in a CRIS system that are considered relevant and can be utilised within the current scope and data model of the OpenAIRE infrastructure. The interoperability use case supported by the Guidelines is the harvesting of data from individual CRIS systems by the OpenAIRE infrastructure. The data will be represented in CERIF XML and will comply with a specialisation of the CERIF XML Schema that defines an OpenAIRE-specific CERIF subset. This is an example of a point-to-point exchange of CRIS information in CERIF XML for a particular application domain. The specialised OpenAIRE CERIF XML Schema is designed so that every XML file that is valid according to it is also valid according to the standard CERIF XML Schema 1.6[1] (when the XML namespace is changed appropriately).

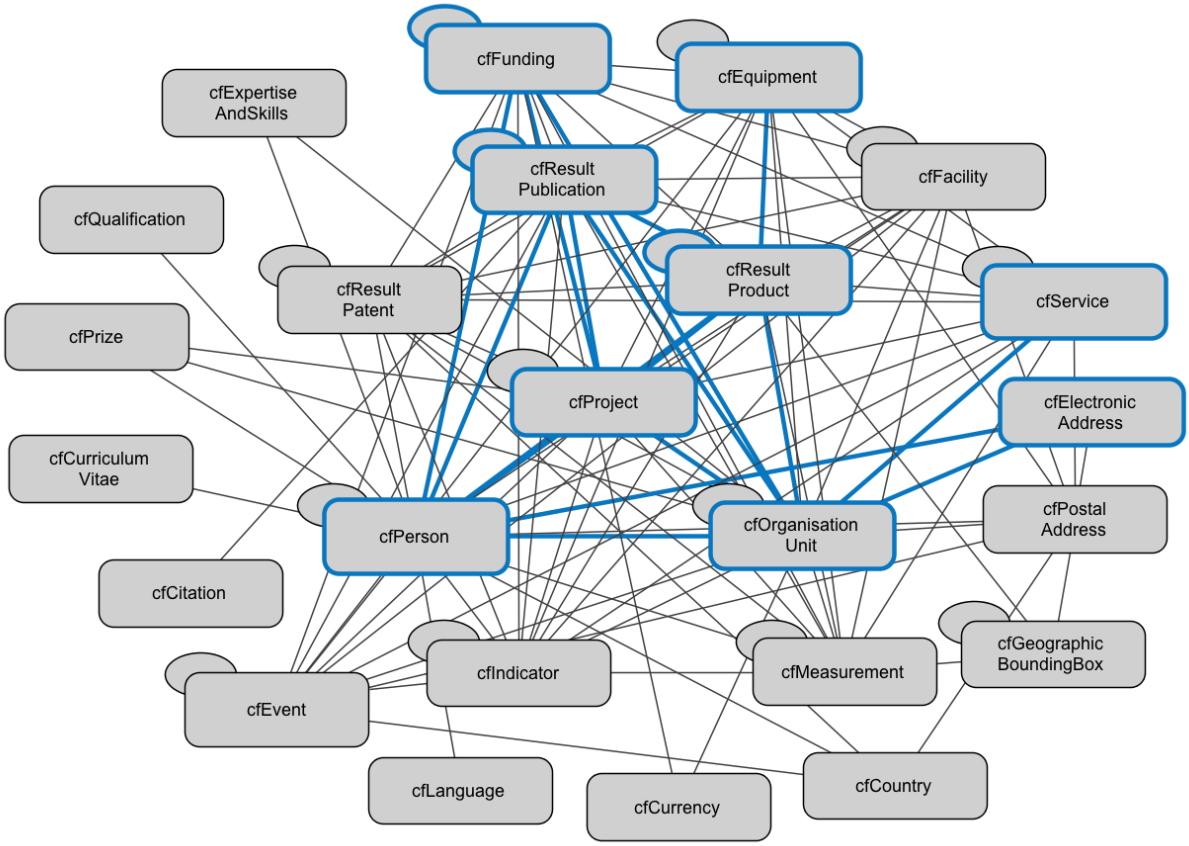

Figure 1: CERIF subset in the current context of OpenAIRE

The Guidelines describe the subset of the CERIF data model and the relevant specialisations that constitute the set of elements for harvesting from CRIS systems through OpenAIRE. The formal model is defined as an OpenAIRE-specific CERIF XML Schema specification which comprises of the following CERIF Research Entities:

[1] See the CERIF 1.6 reference page at http://www.eurocris.org/cerif/feature-tour/cerif-16

- Publication: cfResultPublication *(cfResPubl);*
- Product/Dataset: cfResultProduct *(cfResProd);*
- Person: cfPerson *(cfPers);*
- Organisation: cfOrganisationUnit *(cfOrgUnit);*
- Project: cfProject *(cfProj);*
- Funding: cfFunding *(cfFund);*
- Equipment: cfEquipment *(cfEquip);*
- Service: cfService *(cfSrv).*

The following tables define the CERIF data elements to be utilised for the exchange of data between individual CRIS systems and the OpenAIRE infrastructure. The attributes, classifications, federated identifiers and linking relationships shall be serialized in the order indicated in the tables; the specialized OpenAIRE CERIF XML Schema is more prescriptive in this regard than general CERIF 1.6 XML.
The exclusive use of the defined data elements and vocabularies is mandatory: no other data elements and vocabularies can be used in the CERIF XML data exposed by CRIS systems to the OpenAIRE infrastructure. The vocabularies as currently applied or listed in the below guideline tables are mostly based on the CERIF 1.5 Semantics. Extensions are possible and will be considered for inclusion in future versions of these Guidelines.

2.1 Publication *(cfResPubl)*	The CERIF entity cfResultPublication (*cfResPubl*) is used in the context of OpenAIRE to represent research results that are considered text publications. Metadata about scientific journals or other sources that contain the research results are also represented using the *cfResPubl* entity.	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier *cfResPubl.cfResPublId*		1
Publication Date *cfResPubl.cfResPublDate*		0..1
Journal/Report Number *cfResPubl.cfNum*		0..1
Volume *cfResPubl.cfVol*		0..1
Edition *cfResPubl.cfEdition*		0..1
Issue *cfResPubl.cfIssue*		0..1
Page range start *cfResPubl.cfStartPage*		0..1
Page range end *cfResPubl.cfEndPage*		0..1

Title *cfResPubl.cfTitle*		1
Subtitle *cfResPubl.cfSubTitle*		0..1
Description *cfResPubl.cfAbstr*		1
Subject *cfResPubl.cfKeyw;* *cfResPubl.cfResPubl_Class*	See Section 3.3.	0..N
Language *cfResPubl.cfResPubl_Class*	Use ISO 639-1 (two letter codes), as recommended by CERIF. A sample of the classification is provided in the "OpenAIRE Languages" scheme.	1
Publication Type *cfResPubl.cfResPubl_Class*	The range of allowed types is limited to the following terms: **- Book** **- Book Review** **- Book Chapter Abstract** **- Book Chapter Review** **- Book Series** **- Anthology** **- Monograph** **- Encyclopedia** **- Journal** **- Journal Article** **- Journal Article Abstract** **- Journal Article Review** **- Journal Issue** [2] **- Conference Proceedings** **- Conference Proceedings Article** **- Conference Abstract** **- Conference Poster** **- Letter to Editor** **- Doctoral Thesis** **- Master's Thesis** **- Report** **- Short Communication** **- Commentary** **- Annotation** **- Chapter in Book** **- Working Paper** **- Encyclopedia Entry**	1

[2] Journal issues are expected to be included as types of publications only in cases that they need to be recorded for attribution purposes (e.g. when a journals issue contains exclusively peer-reviewed contributions by a specific project or when a person was a guest editor of a journal issue)

	- Dictionary Entry **- Standard or Policy** from the "Output Types" scheme.	
Open Access Type *cfResPubl.cfResPubl_Class*	The range of allowed types is limited to the following terms: **- Closed Access** **- Embargoed Access** [3] **- Open Access** **- Restricted Access** from the "Open Access Types" scheme.[4]	1
License Type *cfResPubl.cfResPubl_Class*	The range of allowed types is limited to the following terms: **- CC Attribution (CC BY)** **- CC Attribution-NonCommercial (CC BY-NC)** **- CC Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)** **- CC Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)** **- CC Attribution-NoDerivs (CC BY-ND)** **- CC Attribution-ShareAlike (CC BY-SA)** **- CC Zero (CC0)** **- GNU Free Documentation License (GFDL)** from the "License Types" scheme.	0..1
Federated Identifiers *cfResPubl.cfFedId.cfFedId* (where the type of identifier is given through *cfResPubl.cfFedId.cfFedId_Class*)	The range of allowed identifier types is limited to the following terms: **- DOI** **- Handle** **- PMCID** **- ISI-Number** **- SCP-Number** **- ISSN** **- ISBN** **- URL** **- URN** from the "Identifier Types" scheme.	0..N
Relationship(s) with	**Applicable Vocabularies**	**Multiplicity**
Publication (recursive)	The range of allowed relationship types is	0..1

[3] In the case of **Embargoed Access**, the *cfEndDate* of the classification specifies the embargo end date for the output. If the publication becomes open after the embargo period, this can be expressed by adding an **Open Access** classification with *cfStartDate* equal to the embargo date.

[4] Translated from the info:eu-repo-Access-Terms vocabulary (http://purl.org/REP/standards/info-eu-repo#info-eu-repo-AccessRights)

cfResPubl.cfResPubl_ResPubl	limited to the term **- Part** from the "Inter-Publication Relations" scheme.[5]	
Product / Dataset *cfResPubl.cfResPubl_ResProd*	The range of allowed relationship types is limited to the term **- Reference** from the "Inter-Output Relations" scheme.	0..N
Person *cfResPubl.cfPers_ResPubl*	The range of allowed roles is limited to the following terms: **- Author [#]** [6] **- Editor [#]** **- Publisher [#]** from the "Person Output Contributions" scheme.	0..N
Organisation *cfResPubl.cfOrgUnit_ResPubl*	The range of allowed roles is limited to the following terms: **- Author [#]** (for corporate authors) **- Author Institution** **- Editor [#]** (for corporate editors) **- Editor Institution** **- Publisher [#]** from the "Organisation Output Roles" scheme.	0..N
Project *cfResPubl.cfProj_ResPubl*	The range of allowed relationship types is limited to the term **- Originator** from the "Project Output Roles" scheme. I.e. Publication has originator Project.	0..N

[5] Articles shall be related with the journal they appear in using the *cfResPubl_ResPubl* link entity with the "Part" classification term (*cfClassId=eda28bc1-34c5-11e1-b86c-0800200c9a66*) with the direction from the article (*cfResPublId1*) to the host journal (*cfResPublId2*). Similarly, a conference proceedings article connects to the containing conference proceedings, a chapter in a book to the containing book, or a book to its book series.

[6] The **[#]** symbol indicates that the *cfOrder* attribute of the CERIF linking entity can be used to record the order of persons or organisations in the specific role with the research output in the cases where this is a requirement.

2.2 Product / Dataset *(cfResProd)*	The CERIF entity cfResultProduct (*cfResProd*) is used in the context of OpenAIRE to represent research results that are classified as datasets. Datasets are linked with publications, with funded projects, with persons and organisations, and with equipment.	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier *cfResProd.cfResProdId*		1
Name *cfResProd.cfName*		1
Description *cfResProd.cfDescr*		1
Subject *cfResProd.cfKeyw;* *cfResProd.cfResProd_Class*	See Section 3.3.	0..N
Language *cfResProd.cfResProd_Class*	Use ISO 639-1 (two letter codes), as recommended by CERIF. A sample of the classification is provided in the "OpenAIRE Languages" scheme.	0..N
Product Type (OpenAIRE) *cfResProd.cfResProd_Class*	The range of allowed types is limited to the following terms: **- Audiovisual** **- Collection** **- Dataset** **- Image** **- Interactive Resource** **- Model** **- Physical Object** **- Software** **- Sound** **- Text** **- Workflow** from the "OpenAIRE Product Types" scheme (adopted from the OpenAIRE Guidelines for Data Archives[7]).	1

[7] The **Event** and **Service** types are not supported as products of research in this version of the Guidelines. An appropriate representation for these types is foreseen for inclusion in a future revision of the Guidelines.

Product Type (CERIF) *cfResProd.cfResProd_Class*	CERIF defines research products at a more generic level and also comprises a range of product type terms: **- Research data set or database** **- Musical Composition** **- Software** **- Website content** from the "Output Types" scheme. *Note: It is recommended that this classification is specified where it is applicable. Where specified, it shall be compatible with the OpenAIRE Product Type.*	0..1
Open Access Type *cfResProd.cfResProd_Class*	The range of allowed types is limited to the following terms: **- Closed Access** **- Embargoed Access** **- Open Access** **- Restricted Access** from the "Open Access Types" scheme. Similar considerations to the Open Access Type for publications apply here.	1
License Type *cfResProd.cfResProd_Class*	The range of allowed types is limited to the following terms: **- CC Attribution (CC BY)** **- CC Attribution-NonCommercial (CC BY-NC)** **- CC Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)** **- CC Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)** **- CC Attribution-NoDerivs (CC BY-ND)** **- CC Attribution-ShareAlike (CC BY-SA)** **- CC Zero (CC0)** **- European Union Public Licence (EUPL)** **- GNU Free Documentation License (GFDL)** **- GNU General Public License (GPL)** **- GNU Lesser General Public License (LGPL)** **- ODC Attribution (ODC BY)** **- ODC Open Database License (ODC ODbL)** **- ODC Public Domain Dedication and License (ODC PDDL)**, from the "License Types" scheme.	0..1
Federated Identifiers *cfResProd.cfFedId.cfFedId* (where the type of identifier is given through	The range of allowed identifier types is limited to the following terms: **- ARK**	0..N

cfResProd.cfFedId.cfFedId_Class)	**- DOI** **- Handle** **- PURL** **- URN** **- URL** from the "Identifier Types" scheme.	
Relationship(s) with	**Applicable Vocabularies**	**Multiplicity**
Publication *cfResProd.cfResPubl_ResProd*	The range of allowed relationship types is limited to the term **- Reference** from the "Inter-Output Relations" scheme.	0..N
Product / Dataset (recursive) *cfResProd.cfResProd_ResProd*	The range of allowed relationship types is limited to the following terms: **- Citation** **- Derivation** **- Supplement** **- Continuation** **- Metadata** **- Version** **- Part** **- Reference** **- Documentation** **- Compilation** **- Variant** **- Identical** from the "Inter-Product Relations" scheme.	0..N
Person *cfResProd.cfPers_ResProd*	The range of allowed relationship types is limited to the following terms: **- Creator [#]** **- Publisher [#]** from the "Person Output Contributions" scheme	0..N
Organisation *cfResProd.cfOrgUnit_ResProd*	The range of allowed relationship types is limited to the following terms: **- Creator [#]** **- Creator Institution** **- Publisher [#]** from the "Organisation Output Roles" scheme.	0..N
Project *cfResProd.cfProj_ResProd*	The range of allowed relationship types is limited to the term **- Originator** from the "Project Output Roles" scheme. E.g. Dataset has originator Project.	0..N

Equipment *cfResProd.cfResProd_Equip*	The range of allowed relationship types is limited to the term **- Generation** from the "Infrastructure Output Relations" scheme.	0..N

2.3 Person *(cfPers)*	The CERIF entity cfPerson (*cfPers*) is used in the context of OpenAIRE to represent persons that are related to publications (authors etc.), datasets (creators etc.) or projects (contact person for organisation in project).	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier *cfPers.cfPersId*		1
Gender *cfPers.cfGender*	The following standard CERIF codes apply: **m** – Male **f** – Female **u** – Unknown (equivalent to not stating the gender)	0..1
First name(s), Family name(s) and Infix name(s) *cfPers.cfPersName_Pers.cfFirstNames* *cfPers.cfPersName_Pers.cfFamilyNames* *cfPers.cfPersName_Pers.cfOtherNames*	The range of allowed name types is limited to the following terms: **- Presented Name** **- Short Name** **- Passport Name** from the "Person Names" scheme.	1..N
Nationality *cfPers.cfPers_Class*	ISO 3166-1 standard list of country codes. A sample of the classification is provided in the "OpenAIRE Countries" scheme.	0..N
Federated Identifiers *cfPers.cfFedId.cfFedId* (where the type of identifier is given through *cfPers.cfFedId.cfFedId_Class*)	The range of allowed identifier types is limited to the following terms: **- ORCID** **- ResearcherID** **- ScopusAuthorID** **- STAFFID** **- DNR** **- ISNI** from the "Identifier Types" scheme.	0..N
Contact *cfPers.cfPers_EAddr.cfEAddr.cfURI*	The range of allowed relationship types is limited to the following terms: **- Email**	0..N

	- Fax **- Phone** from the "Electronic Address Types" scheme.	
Relationship(s) with	**Applicable Vocabularies**	**Multiplicity**
Publication *cfPers.cfPers_ResPubl*	The range of allowed relationship types is limited to the following terms: **- Author** [#] **- Editor** [#] **- Publisher** [#] from the "Person Output Contributions" scheme.	0..N
Product / Dataset *cfPers.cfPers_ResProd*	The range of allowed relationship types is limited to the following terms: **- Creator** [#] **- Publisher** [#] from the "Organisation Output Roles" scheme.	0..N
Organisation *cfPers.cfPers_OrgUnit*	The term **- Affiliation** from the "Person Organisation Roles" scheme shall be used.	0..N
Project *cfPers.cfProj_Pers*	The range of allowed relationship types is limited to the following terms: **- Organisation Contact In Project** **- Principal Investigator** from the "OpenAIRE Person Organisation Project Relationships" scheme.	0..N
Funding *cfPers.cfPers_Fund*	The range of allowed relationship types is limited to the following terms: **- Applicant** **- Contact** from the "Person Funding Roles" scheme.	0..N

2.4 Organisation *(cfOrgUnit)*	The CERIF entity cfOrganisationUnit (*cfOrgUnit*) is used in the context of OpenAIRE to represent research performing organizations producing research results and/or involved in funded projects (e.g. coordinators, participants), or funder organisations.	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier		1

cfOrgUnit.cfOrgUnitId		
Short name *cfOrgUnit.cfAcro*		0..1
Legal name *cfOrgUnit.cfName*		1
Organisation type *cfOrgUnit.cfOrgUnit_Class*	The range of allowed types is limited to the following terms: **- Higher Education** **- Private non-profit** **- Company** **- Government** **- SME** **- Intergovernmental** **- Research Institute** from the "Organisation Types" scheme.	0..1
Country *cfOrgUnit.cfOrgUnit_Class*	Use ISO 3166-1 standard list of country codes. A sample of the classification is provided in the "OpenAIRE Countries" scheme.	0..1
Organisation Seat Location *cfOrgUnit.cfOrgUnit_Class*	The range of allowed values is limited to the NUTS classification (http://ec.europa.eu/eurostat/web/nuts/overview), a sample of which is provided as the "OpenAIRE Geographical Regions" scheme.	0..N
Federated Identifiers *cfOrgUnit.cfFedId.cfFedId* (where the type of identifier is given through *cfOrgUnit.cfFedId.cfFedId_Class*)	The range of allowed identifier types is limited to the following terms: **- FundRef ID** **- INSTID** **- ISNI** **- National Legal Entity Registration Number** **- PIC** **- UKPRN** **- VAT Identification Number** from the "Identifier Types" scheme.	0..N
Organisation Website Address *cfOrgUnit.cfOrgUnit_EAddr.cfEAddr.cfURI*	The term **- Website** from the "Electronic Address Types" scheme shall be used.	0..1
Relationship(s) with	**Applicable Vocabularies**	**Multiplicity**
Publication *cfOrgUnit.cfOrgUnit_ResPubl*	The range of allowed relationship types is limited to the following terms: **- Author [#]**	0..N

	- Author Institution **- Editor [#]** **- Editor Institution** **- Publisher [#]** from the "Organisation Output Roles" scheme.	
Product / Dataset *cfOrgUnit.cfOrgUnit_ResProd*	The range of allowed relationship types is limited to the following terms: **- Creator [#]** **- Creator Institution** **- Publisher [#]** from the "Organisation Output Roles" scheme.	0..N
Person *cfOrgUnit.cfPers_OrgUnit*	The term **- Affiliation** from the "Person Organisation Roles" scheme shall be used.	0..N
Project *cfOrgUnit.cfProj_OrgUnit*	The range of allowed relationship types is limited to the following terms: **- Coordinator** **- Partner** **- Contractor** **- Funder** **- In kind contributor** **- Applicant** from the "Organisation Project Engagements" scheme.	0..N
Funding *cfOrgUnit.cfOrgUnit_Fund*	The range of allowed relationship types is limited to the following terms: **- Manager** **- Contributor** **- Contact** **- Applicant** **- Issuer** **- Responsible** **- Financier** **- Funder** from the "Organisation Funding Roles" scheme.	0..N
Service *cfOrgUnit.cfOrgUnit_Srv*	The term **- Owner** from the "Organisation Research Infrastructure Roles" scheme shall be used.	0..1

2.5 Project *(cfProj)*	The CERIF entity cfProject (*cfProj*) in the context of OpenAIRE is used to represent funded projects.	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier *cfProj.cfProjId*		1
Start Date *cfProj.cfStartDate*		1
End Date *cfProj.cfEndDate*		1
Acronym *cfProj.cfAcro*		0..1
Title *cfProj.cfTitle*		1
Abstract *cfProj.cfAbstr*		0..1
Subject *cfProj.cfKeyw;* *cfProj.cfProj_Class*	See Section 3.3.	0..N
Open Access Requirements *cfProj.cfProj_Class*	The range of allowed values is limited to the following terms: **- OA mandated** **- OA not mandated** from the "OpenAIRE Open Access Requirements" scheme. *Note: The vocabulary term list is expected to be extended in the future through feedback regarding Open Access mandates internationally.*	0..1
Federated Identifiers *cfProj.cfFedId.cfFedId* (where the type of identifier is given through *cfProj.cfFedId.cfFedId_Class*)	The range of allowed types is limited to the following terms: **- Project Reference** **- URL** from the "Identifier Types" scheme. If URL is used, the term **- Website** from the "Electronic Address Types" scheme shall be applied in addition.	0..N
Relationship(s) with	**Applicable Vocabularies**	**Multiplicity**
Publication *cfProj.Proj_ResPubl*	The term **- Originator**	0..N

	from the "Project Output Roles" scheme shall be used. I.e., Publication has originator Project.	
Product / Dataset *cfProj.cfProj_ResProd*	The term **- Originator** from the "Project Output Roles" scheme shall be used. I.e., Dataset has originator Project.	0..N
Person *cfProj.Proj_Pers*	The range of allowed relationship types is limited to the following terms: **- Organisation Contact In Project** **- Principal Investigator** from the "OpenAIRE Person Organisation Project Relationships" scheme.	0..N
Organisation *cfProj.cfProj_OrgUnit*	The range of allowed relationship types is limited to the following terms: **- Coordinator** **- Partner** **- Contractor** **- Funder** **- In kind contributor** **- Applicant** from the "Organisation Project Engagements" scheme.	1..N
Funding *cfProj.Proj_Fund*	The range of allowed relationship types is limited to the following terms: **- Award** **- Grant** **- Contract** from the "Activity Funding Types" scheme.	0..N

2.6 Funding *(cfFund)*	The CERIF entity cfFunding (*cfFund*) in the context of OpenAIRE is used to represent funding programmes (e.g. EU funded or national programmes) including their structures (using the recursive *cfFund_Fund* linking entity). Funding programmes are linked with projects and organisations (funders).	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier *cfFund.cfFundId*		1
Start date *cfFund.cfStartDate*		0..1

End date *cfFund.cfEndDate*		0..1
Acronym *cfFund.cfAcro*		0..1
Amount *cfFund.cfAmount* *cfFund.cfAmount.cfCurrCode*	Use ISO 4217 standard codes for currencies, as recommended by CERIF. A sample of the classification is provided in the "OpenAIRE Currencies" scheme.	0..1
Name *cfFund.cfName*		1
Description *cfFund.cfDescr*		0..1
Subject *cfFund.cfKeyw;* *cfFund.cfFund_Class*	See Section 3.3.	0..N
Funding Type *cfFund.cfFund_Class*	The range of allowed types is limited to the following terms: **- Funding Programme** **- Call** **- Tender** **- Gift** **- Internal Funding** from the "Funding Source Types" scheme.	0..1
Open Access Requirements *cfFund.cfFund_Class*	The range of allowed types is limited to the following terms: **- OA mandated** **- OA not mandated** from the "OpenAIRE Open Access Requirements" scheme. *Note: The vocabulary is expected to be extended in the future through feedback regarding particular Open Access mandates internationally.*	0..1
Federated Identifiers *cfFund.cfFedId.cfFedId* (where the type of identifier is given through *cfFund.cfFedId.cfFedId_Class*)	The range of allowed identifier types is limited to the following terms: **- Grant Reference** **- URL** from the "Identifier Types" scheme. If URL is used, the term **- Website** from the "Electronic Address Types" scheme shall be applied in addition.	0..N

Relationship(s) with	Applicable Vocabularies	Multiplicity
Person *cfFund.cfPers_Fund*	The range of allowed relationship types is limited to the following terms: **- Applicant** **- Contact** from the "Person Funding Roles" scheme.	0..N
Organisation *cfFund.cfOrgUnit_Fund*	The range of allowed relationship types is limited to the following terms: **- Manager** **- Contributor** **- Contact** **- Applicant** **- Issuer** **- Responsible** **- Financier** **- Funder** from the "Organisation Funding Roles" scheme.	0..N
Project *cfFund.cfProj_Fund*	The range of allowed relationship types is limited to the following terms: **- Award** **- Grant** **- Contract** from the "Activity Funding Types" scheme.	0..N
Funding (recursive) *cfFund.cfFund_Fund*	The term **- Part** from the "Inter-Funding Relations" scheme shall be used.	0..N

2.7 Equipment *(cfEquip)*	The CERIF entity cfEquipment (*cfEquip*) is used in the context of OpenAIRE to represent equipment/devices that are used for the generation of data sets.	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier *cfEquip.cfEquipId*		1
Acronym *cfEquip.cfAcro*		0..1
Name *cfEquip.cfName*		1
Description		0..1

cfEquip.cfDescr		
Federated Identifiers *cfEquip.cfFedId.cfFedId* (where the type of identifier is given through *cfEquip.cfFedId.cfFedId_Class*)	The term **- Institution assigned unique identifier** (such as an asset inventory number) from the "Identifier Types" scheme shall be used.	0..N
Relationship(s) with	**Applicable Vocabularies**	**Multiplicity**
Dataset *cfEquip.cfResProd_Equip*	The term **- Generation** from the "Infrastructure Output Relations" scheme shall be used.	1..N

2.8 Service *(cfSrv)*	The CERIF entity cfService (*cfSrv*) in the context of OpenAIRE can be used to provide information about the CERIF-compliant CRIS that exposes data to OpenAIRE in CERIF XML i.e., the OpenAIRE data source. Each such data source is represented by exactly one *cfSrv* entity. The service is linked with the organisation on whose behalf it runs.	
Attributes	**Applicable Vocabularies**	**Multiplicity**
Internal Identifier *cfSrv.cfSrvId*		1
Name *cfSrv.cfName*		1
Description *cfSrv.cfDescr*		0..1
Service Type *cfSrv.cfSrv_Class*	The type of the service is established using the term - **OpenAIRE CRIS v.1.0** from the "OpenAIRE Service Types" scheme.	1
Website Address *cfSrv.cfFedId.cfFedId*	The term **- URL** from the "Identifier Types" scheme and the term **- Website** from the "Electronic Address Types" scheme shall be used together.	0..1
OAI-PMH Base URL	The term	0..1

cfSrv.cfFedId.cfFedId	**- URL** from the "Identifier Types" scheme and the term **- OAI-PMH 2.0 Repository Endpoint** from the "Electronic Address Types" scheme shall be used together.	
Subject Classification Resource Address[8] *cfSrv.cfFedId.cfFedId*	The term **- URL** from the "Identifier Types" scheme and the term **- Subject Classification Resource Address** from the "Electronic Address Types" scheme shall be used together.	1..N
Relationship(s) with	**Applicable Vocabularies**	**Multiplicity**
Organisation *cfSrv.cfOrgUnit_Srv*	The term: **- Owner** from the "Organisation Research Infrastructure Roles" scheme shall be used.	1

[8] The referenced resource shall contain the controlled vocabularies (classification schemes) used for expressing the subject classifications of publications, products/datasets, projects and funding records provided by the service. See Section 3.3 for details.

3 Technical implementation guidelines

3.1 Data representation in CERIF XML

The CERIF XML style allowed is the one defined in CERIF 1.6 XML specification. The following rules apply – however we refer to the full specification for the details:

1. The CERIF data must be represented as descendants of a root XML element, *CERIF*. Direct descendants of the *CERIF* elements shall only be the following simple CERIF Entities: cfResultPublication *(cfResPubl),* cfResultProduct *(cfResProd),* cfPerson *(cfPers),* cfOrganisationUnit *(cfOrgUnit),* cfProject *(cfProj)*, cfFunding *(cfFund)*, cfEquipment *(cfEquip)*, cfService *(cfSrv)* and cfElectronicAddress *(cfEAddr)*. Stand-alone multi-lingual or link entities or federated identifiers are not accepted as direct descendants of the *CERIF* element.
2. Each CERIF Research Entity in the CERIF XML file embeds its basic attributes, multilingual attributes, link entities and federated identifiers. CERIF Research Entities themselves must not be embedded within another CERIF entity; they can be direct descendants of exclusively the root CERIF XML element. Only links to other entities may be embedded under CERIF Research entities.
3. Link entities must appear at both ends of a relationship. For example, if *cfProj* A is related to *cfOrgUnit* B, the link entity *cfProj_OrgUnit* must appear embedded in the XML record of both *cfProj* A and *cfOrgUnit* B. Therefore, the XML element of every CERIF Research Entity instance must embed link entities for all the relationships to which the entity instance participates. The only exception to this rule is the electronic address (*cfEAddr*), which does not have the reverse links to the objects that use it.
4. Each link entity contains, among others, the identifiers of a classification (term) that denotes the semantics of the relationship and the classification scheme (vocabulary) to which the term belongs. The detailed specification of the classification scheme and its terms is not part of the OpenAIRE CERIF metadata records that are harvested – these are part of the OpenAIRE Semantics that is provided together with these Guidelines. Subject classifications are provided to OpenAIRE separately by the CRIS systems as described in the definition of the cfService entity (Subject Classification Resource Address). Every classification and classification scheme used in link entities should belong to the set of classifications and classification schemes that constitute the OpenAIRE CERIF Semantics specification. Therefore, every identifier used for specifying semantics of link entities in the CERIF XML exposed by CRIS systems should be among the identifiers (UUIDs) contained in the OpenAIRE CERIF Semantics specification.
5. Referential integrity constraints for all relationships among entities should be satisfied in the CERIF XML data provided by the CRIS system. Therefore, it is required that all CERIF objects referenced in the linking relationships in the CERIF XML data are actually represented in the data provided by the same CRIS system. For example, consider the case of a relationship between *cfOrgUnit* A and *cfProj* B that is included in the source CRIS system. To accomplish this, the CERIF XML data exported by the CRIS system must contain:
   a. An XML record for *cfOrgUnit* A. This XML record must contain, as a nested XML element, the link entity *cfProj_OrgUnit*.
   b. An XML record for *cfProj* B. This XML record must contain, as a nested XML element, the link entity *cfProj_OrgUnit*.

It is worth noting that the two aforementioned XML records may be contained in distinct sets of XML records exported by the CRIS system through separate OAI-PMH sets (see Section 3.4.2).

3.2 CERIF Semantic Layer link entities implementation in OpenAIRE CERIF XML

The applicable vocabularies and terms are identified by UUIDs in the CERIF *cfClassScheme.cfClassSchemeId* and *cfClass.cfClassId* attributes, respectively. Those specific vocabularies that must be used by CRIS systems harvested by OpenAIRE are specified in the tables of Section 2.
For the specific case of external vocabularies that are used by the OpenAIRE guidelines (ISO 3166-1 for countries, ISO 639-1 for languages, ISO 4217 for currencies, and NUTS for geographical regions) the *cfClass.cfClassId* attribute contains the codes of the terms in the respective standard (e.g. "GB" for the United Kingdom in ISO 3166-1), while the *cfClass.cfTerm* attribute contains the human readable labels of the corresponding value (e.g. "United Kingdom (the)" for the United Kingdom in ISO 3166-1). UUIDs are not used as *cfClassId* values for these external vocabularies. The associated CERIF semantics file gives a few sample entries.

3.3 Subject Classification

These Guidelines allow the subjects of publications, products/datasets, projects and funding to be expressed in two ways:

1. Using free-text keywords in the *cfKeyw* properties. Multiple keywords in the same language should be expressed as a semi-colon separated list in one *cfKeyw* value.
2. Using a controlled vocabulary (classification scheme).

These Guidelines neither prescribe nor endorse a particular subject classification scheme. The Guidelines do, however, specify the way in which the subject classification scheme used by the CRIS is declared: through the Subject Classification Resource Address in the Service entity (see Section 2.8). The subject classification scheme shall be expressed in CERIF XML 1.6.
The URL of the CERIF XML representation of the subject classification scheme can point either to a static file available at a public web location, or to a file that is produced by the CRIS itself (generated on-demand or cached).

3.4 OAI-PMH for Harvesting

OpenAIRE uses the OAI-PMH v2.0 protocol for harvesting metadata from CRIS systems.

3.4.1 Metadata Format

OpenAIRE expects metadata from CRIS systems to be encoded in the CERIF XML metadata format, as specialised for OpenAIRE in the accompanying XML Schema and semantic vocabularies and described in the previous section. The following metadata prefix should be used: **cerif_openaire**.

3.4.2 OpenAIRE OAI-PMH Sets

For harvesting the records relevant to OpenAIRE, the use of specific OAI-PMH sets at the local CRIS system is mandatory. The description and required characteristics of the sets are provided in the following table:

Description	Required characteristics
The entire graph of CERIF entities in the source CRIS system of relevance to OpenAIRE. The semantic vocabularies are not included: they are rather specified by (a) the OpenAIRE Semantics files that accompany these Guidelines, and (b) the subject classification files whose locations are specified in the cfService description of the OpenAIRE data source.	**setName:** OpenAIRE_CRIS **setSpec:** openaire_cris
The list of CERIF XML records for persons with embedded multi-lingual entities, federated identifiers and link entities and for associated electronic addresses.	**setName:** OpenAIRE_CRIS_persons **setSpec:** openaire_cris_persons
The list of CERIF XML records for projects with embedded multi-lingual entities, federated identifiers and link entities.	**setName:** OpenAIRE_CRIS_projects **setSpec:** openaire_cris_projects
The list of CERIF XML records for organisation units with embedded multi-lingual entities, federated identifiers and link entities.	**setName:** OpenAIRE_CRIS_orgunits **setSpec:** openaire_cris_orgunits
The list of CERIF XML records for funding with embedded multi-lingual entities, federated identifiers and link entities.	**setName:** OpenAIRE_CRIS_funding **setSpec:** openaire_cris_funding
The list of CERIF XML records for publications with embedded multi-lingual entities, federated identifiers and link entities.	**setName:** OpenAIRE_CRIS_publications **setSpec:** openaire_cris_publications
The list of CERIF XML records for datasets with embedded multi-lingual entities, federated identifiers and link entities and for associated equipment.	**setName:** OpenAIRE_CRIS_datasets **setSpec:** openaire_cris_datasets
The CERIF XML record representing the CRIS as an OpenAIRE data source, with embedded multi-lingual entities and link entities.	**setName:** OpenAIRE_CRIS_services **setSpec:** openaire_cris_services

Referential integrity constraints for all relationships among entities should be satisfied in the CERIF XML data provided by the CRIS system, as mentioned in the "Data representation in CERIF XML" sub-section above. This holds also for the case that entity instances related via link entities are retrieved through different OAI-PMH sets. For example, consider the case of a relationship between *cfOrgUnit* A and *cfProj* B that is included in the source CRIS system. To accomplish this, the CERIF XML data exported by

the CRIS system must contain:

a. An XML record for *cfOrgUnit* A. This XML record must contain, as a nested XML element, the link entity *cfProj_OrgUnit*. The XML record of *cfOrgUnit* A must be available through both sets ***openaire_cris*** and ***openaire_cris_orgunits***.
b. An XML record for *cfProj* B. This XML record must contain, as a nested XML element, the link entity *cfProj_OrgUnit*. The XML record of *cfProj* B must be available through both sets ***openaire_cris*** and ***openaire_cris_projects***.

In case the two entity instances (*cfOrgUnit* A and *cfProj* B) are retrieved via the different sets ***openaire_cris_orgunits*** and ***openaire_cris_projects***, the OAI-PMH service provider – in this case the OpenAIRE infrastructure – should combine and check the information in the two different sets of XML records to validate the source data in terms of referential integrity.

3.4.3 Transmission of CERIF XML as OAI-PMH payload

OAI-PMH is a protocol for exposing information from data providers to clients (service providers). Data provided through OAI-PMH must be encoded in XML and is organised into a sequence of records. The protocol uses the resumption token mechanism to enable control over the flow of data from the data provider towards the service provider, for example, it allows the split of a large chunk of records into fragments of manageable size (e.g. 100 records). This helps avoid overload of both the data and service provider.

Data in CERIF CRIS systems follows a normalised graph structure. Therefore, the transmission of CERIF XML as OAI-PMH payload requires a mechanism of fitting the graph structure into a sequence of records. Each OAI-PMH XML record should represent a single instance of a CERIF Research Entity, which embeds its multi-lingual entities, federated identifiers and link entities. The CERIF XML *cfPers* and *cfOrgUnit* elements are optionally followed by the *cfEAddr* elements that represent the electronic addresses associated with them. The CERIF XML *cfResProd* elements may be followed by any *cfEquip* elements that represent the equipment used to generate the dataset.

3.4.4 Date stamps in CERIF XML records

In OAI-PMH, selective harvesting based on last-update date stamps on records is possible, so that only records that have been modified since the last harvesting are retrieved. Due to considerations regarding not consistent and reliable mechanisms for setting date stamp values in certain source systems, OpenAIRE in the general case 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.

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 when any information contained within the record payload was modified (e.g. entity fields, multilingual fields, federated identifiers, link 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.

For example, we assume a CERIF XML record of type *cfProj*, containing:

1. Entity fields (e.g. *cfProj.cfAcro*);
2. Multilingual fields (e.g. *cfProj.cfTitle*);
3. Federated identifiers (e.g. the project reference in the case of EU FP7 projects);
4. Link entities (e.g. the link entity *cfProj_OrgUnit* denoting relationships of organisations to the project).

Let us consider a CRIS system from which the CERIF XML is exported. Example update cases of a *cfProj* instance and the corresponding date stamp modifications of the *cfProj* CERIF XML record are provided in the following list:

a. *cfProj.cfAcro* is modified. The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the modification.
b. An existing *cfProj.cfTitle* instance is modified, e.g. update of *cfProj.cfTitle*. The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the modification.
c. A new *cfProj.cfTitle* is added to this *cfProj* instance (e.g. the project title in another language). The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the addition.
d. An existing federated identifier instance referring to this *cfProj* instance is modified, e.g. update of the *cfProj.cfFedId.cfStartDate*. The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the modification.
e. An existing federated identifier instance classification referring to this *cfProj* instance is modified, e.g. update of the *cfFedId_Class.cfEndDate*. The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the modification.
f. A new federated identifier is added for this *cfProj* instance. The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the addition.
g. A new *cfProj_OrgUnit* instance is added to the database, linking the *cfProj* instance to an organisation that participates in the project. The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the addition.
h. An existing *cfProj_OrgUnit* link entity instance is modified (e.g. *cfProj_OrgUnit.cfStartDate*). The date stamp of the respective *cfProj* CERIF XML record must be updated to the date/time of the modification.
i. An existing *cfOrgUnit* instance is modified (e.g. *cfOrgUnit.cfAcro*). This particular *cfOrgUnit* instance concerns an organization that is a partner in the project and thus is already connected with the *cfProj* through a *cfProj_OrgUnit* linking entity. In this case, the date stamp of the respective *cfProj* CERIF XML record must NOT be updated to the date/time of the modification.

3.4.5 Deleted records

OpenAIRE does **not** require CRIS systems to provide information about deleted records via OAI-PMH. Therefore, it is acceptable for a CRIS system exposing metadata records to the OpenAIRE infrastructure to provide any of the three levels of support of deleted records, as defined in the OAI-PMH 2.0 specification: "**no**", "**persistent**" or "**transient**". As mandated in the OAI-PMH 2.0 specification, CRIS systems must declare the level of support of deleted records in the *deletedRecord* element of the *Identify* response.

Appendix A: An overview of the examples