Open Access
Infrastructure
for Research in Europe
OpenAIRE

OpenAIRE Guidelines for Data Archive Managers 1.0

June 2013

Contents

Introduction

Aim

OpenAIRE gathers together research output related to European funding streams, with the aim of supporting open science and tracking research impact. This content consists of open access publications and related contextual information such as datasets, supplementary material, and research/project information.

Two other sets of OpenAIRE guidelines exist for repository managers:

- OpenAIRE Guidelines for Literature Repositories 3.0
- OpenAIRE Guidelines for CRIS systems (CERIF standard) – *In draft*

The OpenAIRE Guidelines for Data Archive Managers 1.0 will provide instruction for data archive managers to expose their metadata in a way that is compatible with the OpenAIRE infrastructure. The metadata from data archives should be included in the OpenAIRE information space, and exposed when data are related to an open access publication e.g. a dataset cited by an article.

By implementing the OpenAIRE Guidelines data archive managers are facilitating the creation of enhanced publications and building the stepping-stones for a linked data infrastructure for research.

Exposure and visibility of content from a range of European repositories will be significantly increased when a common and interoperable approach is taken and care to adhere to existing guidelines. OpenAIRE is happy to assist in adherence to these guidelines. This compatibility will lead to future interoperability between research infrastructures, and structured metadata is of benefit to individual data repositories and the scholarly community at large.

DataCite

OpenAIRE has adopted the DataCite Metadata Schema as the basis for harvesting and importing metadata about datasets from data archives.

The core mission of DataCite is to build and maintain a sustainable framework that makes it possible to cite data through the use of persistent identifiers, DOIs.

OpenAIRE shares the goal of the DataCite Metadata Schema - to provide a domain agnostic metadata schema and provide interoperability through a small number of properties - making interoperability possible in the simplest manner possible and as a result keep the technical barriers for implementation as low as possible.

We would like to thank DataCite for their kind support and help making these OpenAIRE Guidelines for Data Archive Managers possible.

What’s different

In this document you will find the needed properties to make your data archive compatible with the OpenAIRE infrastructure. OpenAIRE has adopted the DataCite Metadata Schema version 2.2 with some minor adjustments.

- OpenAIRE accepts other persistent identifier schemes and not only a DOI in *1.1 identifierType*.
- OpenAIRE recommends exporting links to related publications and datasets (i.e.
  12. RelatedIdentifier property and attributes are mandatory when applicable).
- OpenAIRE recommends using the *7. Contributor* property to relate a dataset to funding information.
- DataCite optional properties *8. Date* and *17. Description* are mandatory in OpenAIRE.
- OpenAIRE enforces an encoding scheme on DataCite property *16. Rights*.

Acknowledgements & Contributors

Editors

- Mikael K. Elbæk (Technical University of Denmark)
- Lars Holm Nielsen (CERN, Switzerland)

Experts & Reviewers

- Samuele Kaplun (CERN, Switzerland)
- Paolo Manghi (CNR, Italy)
- Natalia Manola (University of Athens, Greece)
- Jochen Schirrwagen (University of Bielefeld, Germany)
- Birgit Schmidt (University of Goettingen,Germany)
- Mathias Lösch (University of Bielefeld, Germany)
- Frauke Ziedorn (DataCite, Germany)
- Pedro Principe (University of Minho, Portugal)
- Eloy Rodrigues (University of Minho, Portugal)
- Najla Rettberg (University of Goettingen, Germany)
- Jeroen Rombouts (TU Delft, The Netherlands)
- Jane Greenberg, (University of North Carolina at Chapel Hill)

License

Version

1.0	June 2013	Initial document

OpenAIRE application of DataCite Metadata Schema

OpenAIRE builds on the DataCite Metadata Schema v2.2 by making some of the otherwise optional DataCite properties mandatory, as well as enforcing specific encoding schemes on the values of some DataCite properties. The table below details the differences from the DataCite Metadata Schema.

Within OpenAIRE the use of elements and attributes is either:

- **mandatory (M)** = the element must always be present in the metadata record. An empty element is not allowed.
- **mandatory when applicable (MA)** = when the element can be obtained it must be present in the metadata record
- **recommended (R)** = the use of the element is recommended
- **optional (O)** = the element may be used to provide complementary information about the resource

The recommended status is made primarily to encourage users to input certain elements when creating a metadata record to enhance services.

Important: The table below details only differences from the DataCite Metadata Schema v2.2. Please refer to the DataCite Metadata Schema v2.2 for definition of elements and their encoding schemes: http://schema.datacite.org/meta/kernel-2.2/index.html

ID	DataCite-property	Sta-tus	Encoding schemes (if different from DataCite)[1]
1	Identifier	M	DOI (Digital Object Identifier) is preferred. The format should be "10.1234/foo".
1.1	identifierType	M	Unlike DataCite, OpenAIRE allows for DOIs and other types of identifiers.

[1] Only differences to the DataCite Metadata Schema v2.2 are noted here. Please refer to encoding schemes and allowed values in http://schema.datacite.org/meta/kernel-2.2/index.html

			Controlled List: Allowed values: ARK DOI Handle PURL URN URL
2	Creator	M	
2.1	creatorName	M	
2.2	nameIdentifier	R	
2.2.1	nameIdentifierScheme	R	
3	Title	M	
3.1	titleType	O	
4	Publisher	M	
5	PublicationYear	M	
6	Subject	O	
6.1	subjectScheme	O	
7	Contributor	MA /O	OpenAIRE uses this element to allow unique and persistent identification of the funder who has funded wholly or partly the dataset described. This does not exclude also using this element for additional contributors as defined by DataCite Metadata Schema v2.2. The element and sub-elements are only *mandatory when applicable (MA)* for describing funding information. Further use is *optional (O)*. Please refer to the section “Funding information” at the end of the table for a

			complete example of the use of *7. Contributor, 7.1 contributorType, 7.2 contributorName, 7.3 nameIdentifier, 7.3.1 nameIdentifierScheme* to provide funding information.
7.1	contributorType	MA /O	*Controlled List* Allowed values (in-complete): Funder Please see DataCite Metadata Schema v2.2 for all allowed values.
7.2	contributorName	MA /O	Applicable only when contributorType is "Funder": Name of the funding entity. Example for European Commission funded research use "European Commission", or for Wellcome Trust funded research use "Wellcome Trust". Specifically *do not* use the project acronym.
7.3	nameIdentifier	MA /O	Applicable only when contributorType is "Funder": A vocabulary of projects[2] is exposed by OpenAIRE through OAI-PMH, and available for all repository managers. Values will include the project name and projectID. The projectID equals the Grant Agreement identifier, and is defined by the info:eu-repo namespace term grantAgreement[3] The syntax is: `info:eu-`

[2] See http://api.openaire.research-infrastructures.eu:8280/is/mvc/openaireOAI/oai.do?verb=ListRecords&set=projects&metadataPrefix=oaf

[3] See http://purl.org/REP/standards/info-eu-repo#info-eu-repo-GrantAgreementIdentifiers

			`repo/grantAgreement/Funder/FundingProgram/ProjectID/[Jurisdiction]/[ProjectName]/[ProjectAcronym]/` where: • `Funder` refers to the funding origanization (e.g., EC for European Commission, WT for Wellcome Trust) • `FundingProgramme` refers to a specific programme (e.g., FP7) • `ProjectID` refers to a unique identifier in the scope of the funder (and maybe the programme), e.g. a grant agreement number. • `Jurisdiction` refers to the authority granted to a formally constituted legal body (e.g. EU for European Union) • `ProjectName` contains the full name of the project • `ProjectAcronym` contains the project's acronym. For OpenAIRE compatibility, the elements in square brackets are optional. Repositories may choose to use only the old three-part namespace (`info:eu-repo/grantAgreement/Funder/FundingProgram/ProjectID`), or use the extended version with six parts, which is recommended. **Note:** When omitting fields in the extended version, the number of fields must nevertheless be preserved by using "/". A correct example for omitting the the `ProjectName` field would therefore look like this: `EC/FP7/12345/EU//OpenAIREplus`

7.3.1	nameIdentifierScheme	MA /O	Applicable only when contributorType is “Funder”: *Controlled List* Allowed values: info
8	Date	M	*Mandatory* property in OpenAIRE instead of *optional* in DataCite. For encoding scheme, please refer to DataCite Metadata Schema v2.2 for details.
8.1	dateType	M	Use “Issued” for the date the resource is published or distributed. To indicate the end of an embargo period, use “Available”. To indicate the start of an embargo period, use “Accepted”. Further dateTypes may be specified. Please refer to DataCite Metadata Schema v2.2 for details.
9	Language	R	
10	ResourceType	R	
10.1	resourceTypeGeneral	R	
11	AlternateIdentifier	O	
11.1	alternateIdentifierType	O	
12.	RelatedIdentifier	MA	*Mandatory when applicable* property in OpenAIRE instead of optional in DataCite. Please refer to the section “Related publications and datasets information” below for specific details on how to link datasets and publications.
12.1	relatedIdentifierType	MA	
12.2	relationType	MA	*Controlled List* Allowed values (recommended ones,

			please refer to DataCite Metadata Schema v2.2 for a complete list) IsCitedBy (indicates that B includes A in a citation) Cites (indicates that A includes B in a citation) IsReferencedBy (indicates A is used as a source of information by B) References (indicates B is used as a source of information for A) IsPartOf (indicates A is a portion of B; may be used for elements of a series) HasPart (indicates A includes the part B) IsNewVersionOf (indicates A is a new edition of B, where the new edition has been modified or updated) IsPreviousVersionOf (indicates A is a previous edition of B) **Note:** *Cites* and *IsCitedBy* is specifically for when a publication/dataset directly cites another publication/dataset in its references, whereas *References* and *IsReferencedBy* is for when a dataset/publication is used as a source of information without a direct citation. OpenAIRE encourages including minimum one of above listed relation types, but allows usage of all DataCite's relation types.
13	Size	O	
14	Format	O	
15	Version	O	
16	Rights	MA	*Mandatory when applicable* property in OpenAIRE instead of optional in DataCite. Use terms from the info:eu-repo-Access-

			Terms vocabulary[4]. The values are • info:eu-repo/semantics/closedAccess • info:eu-repo/semantics/embargoedAccess • info:eu-repo/semantics/restrictedAccess • info:eu-repo/semantics/openAccess If the material is licensed under a Creative Commons license then you should provide links to applicable Creative Commons licenses, e.g.: http://creativecommons.org/licenses/zero/1.0/ http://creativecommons.org/licenses/by/3.0/
17	Description	MA	*Mandatory when applicable* property in OpenAIRE instead of optional in DataCite.
17.1	descriptionType	MA	

The OpenAIRE Guidelines for Data Archive Managers are built on the DataCite Metadata Schema[5]. The following properties are described in further detail to make full integration into the OpenAIRE information space and allow OpenAIRE to make links between publications and datasets; datasets and funding information.

[4] See http://purl.org/REP/standards/info-eu-repo#info-eu-repo-AccessRights

[5] DataCite Metadata Schema version 2.2: http://schema.datacite.org/meta/kernel-2.2/index.html

Access rights information

OpenAIRE uses the access rights to enable a better user experience by declaring the access rights clear and explicit in the portal. Access rights are specified using the *16. Rights* property. Please see encoding scheme in the section above.

An example:

```
<rights>info:eu-repo/semantics/openAccess</rights>
```

Funding information

One of OpenAIRE's main goals is to link research output to (EC) research funding. The following application of the contributor property allows unique and persistent identification of the funder who has funded wholly or partly the dataset described.

The following properties are mandatory when applicable to provide funding information: 7. *Contributor*; *7.1 contributorType*; *7.2 contributerName*; *7.3 nameIdentifier*; *7.3.1 nameIdentifierScheme*:

An example for linking a research output to the OpenAIREplus FP7 project:

```
<contributor contributorType="Funder">
 <contributorName>
     European Commission
 </contributorName>
 <nameIdentifier nameIdentifierScheme="info">
     info:eu-repo/grantAgreement/EC/FP7/282896
 </nameIdentifier>
</contributor>
```

Related publications and datasets information

OpenAIRE **harvests all datasets** from a data repository, but **exposes only certain datasets** in the OpenAIRE portal. See the section "OpenAIRE OAI Set" below for specific details of which datasets are exposed.

For example, datasets related to publication will be exposed in the OpenAIRE portal. The link between the dataset and publication may be explicit defined, as described in this section below, or automatically inferred by the OpenAIRE infrastructure. If the link is explicit defined, the dataset will be exposed in the OpenAIRE portal **within 1-2 days after harvesting** (a repository is harvested once a week on average). If the link is automatically inferred by the OpenAIRE infrastructure it may take **up to a month after harvesting** before the dataset is exposed in the OpenAIRE portal. It is thus *mandatory when applicable* to provide links to related publications and datasets when these links are available in the repository, and thereby ensure faster exposure of the dataset in the OpenAIRE portal.

Related identifiers

DataCite Metadata Schema allows linking publications and datasets by use of persistent identifiers to uniquely identify the resource being described (A) typically a dataset but not limited to that, and the related resource (B) in the case of OpenAIRE typically a publication or a dataset.

Related publications/datasets must have the following properties: *12. RelatedIdentifier*, *12.1 relatedIdentifierType*, *12.2 relationType*

An example:

```
<relatedIdentifier
relatedIdentifierType="DOI"  relationType="IsCitedBy">
 10.1234/bar
</relatedIdentifier>
```

Embargo date information

For OpenAIRE two main types of dates are relevant. When the data were made available, published or uploaded to a formal database, this is the date the data were *Issued*.

Sometimes data may be embargoed for a period; this information should be managed by the data provider and expressed by exporting an *Available* date to indicate the end of an embargo period and an *Accepted* date to indicate the start of an embargo period.

An example:

```
<dates>
 <date dateType="Issued">2011-12-01</date>
</dates>
```

An embargo example:

```
<dates>
 <date dateType="Accepted">2011-12-01</date>
 <date dateType="Available">2012-12-01</date>
</dates>
```

Use of OAI-PMH

OpenAIRE uses the OAI-PMH v2.0 protocol for harvesting dataset metadata.

Metadata Format

OpenAIRE expects metadata to be encoded in the *DataCite* metadata format (metadataPrefix `oai_datacite`)[6]. For information on how to use the individual DataCite elements, please refer to the section "OpenAIRE application of DataCite Metadata Schema" above.

OpenAIRE OAI Set

For harvesting the records relevant to OpenAIRE, the use of a specific OAI set at the local repository is mandatory. The set must have the following characteristics:

setName	setSpec*
OpenAIRE_data	`openaire_data`

A harvester only uses the* ***setSpec* *request to perform selective harvesting. The letters must be in small caps.*

Set content

The specific content of the `openaire_data` set is to be determined at the local repository. OpenAIRE will **harvest all datasets** from the `openaire_data` set, but will **only expose datasets fulfilling at least one** of the following criteria in the OpenAIRE portal:

- The dataset is outcome of a funded research project identified by a project identifier (see section "Funding information" above).
- The dataset is linked with a publication in the OpenAIRE information space (see section "Related publications and datasets information" above).

Both criteria above may be automatically inferred by the OpenAIRE infrastructure.

Specifically a data repository may insert a dataset without funding information or link to a related publication in the `openaire_data` set. If later a publication harvested from a literature repository links to the dataset, the dataset will be exposed in the OpenAIRE portal. If either the literature repository or the data repository explicitly

[6] See http://schema.datacite.org/oai/oai-1.0/ and http://oai.datacite.org/

links the publication and dataset (see section ”Related publications and datasets information”), the dataset will normally be exposed in the OpenAIRE portal **within 1-2 days after harvesting** (repositories are harvested on average once a week). If the OpenAIRE infrastructure automatically has to infer the link between the publication and the dataset, it may take **up to a month after harvesting** before the dataset is exposed in the OpenAIRE portal.

Futher instructions

Best practices and cases of implementations will continually be updated on the OpenAIRE Guidelines wiki: https://guidelines.openaire.eu

The wiki is also intended platform for having a dialogue with the user communities to discuss the desired development and direction of the guidelines.