OpenAIRE Content Acquisition Policies

Version: 1.0 **Release date:** 5th of October, 2018

Authors: Amelie Becker[3], Aenne Loeden[3], Paolo Manghi[1], Pedro Principe[2], Jochen Schirrwagen[3]

Contributors: Claudio Atzori[1], Miriam Baglioni[1], Alessia Bardi[1], Andreas Czerniak[3], Natalia Manola[4], Eloy Rodrigues[2]

[1] Institute of information science and technologies - CNR, Pisa, Italy
[2] University of Minho - Minho, Portugal
[3] Bielefeld University - Bielefeld, Germany
[4] Athena Research & Innovation Center - Athens, Greece

Summary This document defines the conditions under which metadata of scientific products collected from content providers in OpenAIRE will be considered for inclusion in the OpenAIRE information space. Policies specify which typologies of objects are mapped into which OpenAIRE entities (literature, dataset, software, other research products) and which are the minimal quality conditions under which metadata can be accepted.

Table of content

1 Rationale **2**

2. Aggregation policies by type of product **2**

- 2.1 Literature, Datasets, Software, other research products 2
- 2.2 Accession numbers 2
- 2.3 Full-text of scientific literature 3

3. Semantic mapping: from metadata types to OpenAIRE entities **3**

1 Rationale

The OpenAIRE service infrastructure harvests metadata about scholarly communication products (literature, datasets, software, and other research products) and links between such products from a range of institutional or subject repositories, national and institutional research information portals, aggregators, e-journals, data repositories, and software repositories. In addition, it infers links between literature and such products via advanced text and data mining techniques (TDM). The resulting information graph (i.e. interlinked sets of objects) is intended to favour monitoring of open science and open science publishing workflows (e.g. science reproducibility and transparent assessment).

- **Coverage** OpenAIRE will actively pursue harvesting content from European but also non-European repositories;
- **Reproducibility** The OpenAIRE graph aims at linking scientific literature, namely the narration of scientific motivation and process, with all products used or resulting in the relative research activity
- **Monitoring** The OpenAIRE graph links research products with the funders and projects resulting from their grants
- **Research communities** The OpenAIRE graph links research products with the communities for which they are relevant, in order to provide a (multi-)community-view of the scholarly output
- **Quality** Data sources and repositories are quality-controlled: their metadata respects the OpenAIRE guidelines: https://guidelines.openaire.eu and their import in OpenAIRE is curated by OpenAIRE data curators. Click here to see the repositories that OpenAIRE currently harvests from.
- **Terms of Agreement for content providers** Data source managers read and accept the OpenAIRE Terms of Agreement in order for OpenAIRE to re-use their metadata and Open Access full-text under specific consent, warranties, and license.

2. Aggregation policies by type of product

2.1 Literature, Datasets, Software, other research products

OpenAIRE accepts the metadata records of all scientific products whose structure respect the model and semantics as expressed by the OpenAIRE guidelines. This means that both Open Access and non-Open Access material will be included and links to other products will be resolved where this is possible (i.e. the provided PIDs have a resolver).

2.2 Accession numbers

Datasets with accession numbers (database entries) are not included as OpenAIRE datasets but, when a relationship to product exists, are included as properties of the related products. More specifically, they are included as values of the property *externalReference* of product metadata; *externalReference* includes a *URL to the splash page*, the *target web site name*, the *ID* and an *ID type* (e.g. PDB).

2.3 Full-text of scientific literature

OpenAIRE collects Open Access literature product files whenever these are accessible from the URL provided in the metadata record. The literature full-text is used for text-mining purposes. End-users willing to access, download, and read the actual files will not be able to do so from OpenAIRE, but will be forwarded to the original source of deposition. For further information on the use of full-texts, please view OpenAIRE's Terms of Agreement with content providers.

3. Semantic mapping: from metadata types to OpenAIRE entities

OpenAIRE services collect metadata about four typologies of products: literature, datasets, software, and "other research products" (ORPs). Metadata can be collected from four main categories of repositories: *literature repositories* (including institutional/thematic repositories, publishers, and catalogues), *data repositories*, *software repositories*, and *ORP repositories*. As things stands there is no one-to-one relationships between a type of repository and the products it contains, e.g. literature repositories may indeed also contain datasets, software, and ORPs. Accordingly, the aggregation process needs to classify the products collected from a repository in order to assign them to the correct entity class in OpenAIRE. The distribution rules are illustrated in Table 1, which follow vocabularies in the OpenAIRE guidelines and Version 4.0 of DataCite. Please note that such mappings may be modified over time to reflect the general preferences and requirements of the OpenAIRE user community.

Table 1. Product types and content provider guidelines classes

	Literature type	**Dataset type**	**Software type**	**Other research product type** (any product that is not of type literature, dataset, or software)
Guidelines for literature repositories (v4.0) Includes: publishers, journals, institutional repositories, aggregators, catalogues	Resource type different from the ones associated to Dataset, Software, and Other products	Resource type indicating datasets, image, video, audio	Resource type indicating software	Resource type indicating other research products (e.g. "Service", "Interactive Resource", "Other" etc.)
Guidelines for data repositories	Resource type indicating papers	Resource type different from the	Resource type indicating software	Resource type indicating other

Includes: data repositories, aggregators	(based on repository specific vocabularies)	ones associated to Literature, Software and Other products		research products (e.g. “Service”, “Interactive Resource”, “Other” etc.)
Guidelines for Software repositories: Software repositories	none	none	All records	none
Guidelines for Other research products repositories: other product repositories	none	none	none	All records

Research products type and OpenAIRE entities The detailed mappings between input records and the target OpenAIRE entities are specified below. Independently from the category of input repository, the aggregation process identifies for each input record a term in a common vocabulary. Each term of this vocabulary is then associated to one of the four OpenAIRE entities literature, dataset, software, and ORP as described in the following tables.

Remark: such associations may be modified over time to reflect the general preferences and requirements of the OpenAIRE user community.

Table 2. Research product “Literature”

OpenAIRE Encoding	Term
0001	Article
0002	Book
0004	Conference object
0005	Contribution for newspaper or weekly magazine
0006	Doctoral thesis
0007	Master thesis
0008	Bachelor thesis
0009	External research report
0011	Internal report
0012	Newsletter
0013	Part of book or chapter of book
0014	Research
0015	Review
0016	Preprint
0017	Report
0019	Patent
0031	Data Paper

0032	Software Paper
0034	Project deliverable
0035	Project milestone
0036	Project proposal
0038	Other literature type

Table 3. Research product “Dataset”

OpenAIRE Encoding	Term
0021	Dataset
0024	Film
0025	Image
0030	Sound
0033	Audiovisual
0037	Clinical Trial
0039	Other dataset type

Table 4. Research product “Software”

OpenAIRE Encoding	Term
0029	Software
0040	Other software type

Table 5. Research product “Other research product”

OpenAIRE Encoding	Term
0010	Lecture
0018	Annotation
0023	Event
0026	Interactive resource
0027	Model
0028	Physical object
0020	Other ORP type