Istituto di Scienza e Tecnologie
dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Technical Reports

Data Model description of the OpenAIRE Research Graph

Sandro Fabrizio La Bruzzo, ISTI CNR, Pisa, Italy

Michele Artini, ISTI-CNR, Pisa, Italy

Claudio Atzori, ISTI-CNR, Pisa, Italy

Alessia Bardi, ISTI-CNR, Pisa, Italy

Miriam Baglioni, ISTI-CNR, Pisa, Italy

Michele De Bonis, ISTI-CNR, Pisa, Italy

Andrea Mannocci, ISTI-CNR, Pisa, Italy

Paolo Manghi, ISTI-CNR, Pisa, Italy

Gina Pavone, ISTI-CNR, Pisa, Italy

ISTI-TR-2022/031

Data Model description of the OpenAIRE Research Graph

La Bruzzo S.F.; Artini M.; Atzori C.; Bardi A.; Baglioni M.; De Bonis M.; Mannocci A.; Manghi P.; Pavone G.

ISTI-TR-2022/031

The OpenAIRE Graph (formerly known as the OpenAIRE Research Graph) is one of the largest open scholarly record collections worldwide, key to fostering Open Science and establishing its practices in daily research activities. Conceived as a public and transparent good, populated out of data sources trusted by scientists, the Graph aims at bringing discovery, monitoring, and assessment of science back into the hands of the scientific community. Imagine a vast collection of research products all linked together, contextualized, and openly available. For the past years, OpenAIRE has been working to gather this valuable record. It is a massive collection of metadata and links between scientific products such as articles, datasets, software, and other research products, entities like organizations, funders, funding streams, projects, communities, and data sources. This technical Report describes the public data model adopted by the OpenAIRE Graph.

Keywords: Data Model, Research Graph, OpenAIRE.

Citation
La Bruzzo S.F.; Artini M.; Atzori C.; Bardi A.; Baglioni M.; De Bonis M.; Mannocci A.; Manghi P.; Pavone G. *Data Model description of the OpenAIRE Research Graph*. ISTI Technical Reports 2022/031. DOI: 10.32079/ISTI-TR-2022/031.

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

Data Model description of the OpenAIRE Research Graph

Sandro Fabrizio La Bruzzo (ISTI CNR), Michele Artini (ISTI CNR), Claudio Atzori (ISTI CNR), Alessia Bardi (ISTI CNR), Miriam Baglioni (ISTI CNR), Michele De Bonis (ISTI CNR), Andrea Mannocci (ISTI CNR), Paolo Manghi (ISTI CNR), Gina Pavone (ISTI CNR)

The OpenAIRE Graph (formerly known as the OpenAIRE Research Graph) is one of the largest open scholarly record collections worldwide, key in fostering Open Science and establishing its practices in the daily research activities. Conceived as a public and transparent good, populated out of data sources trusted by scientists, the Graph aims at bringing discovery, monitoring, and assessment of science back in the hands of the scientific community.

Imagine a vast collection of research products all linked together, contextualised and openly available. For the past years OpenAIRE has been working to gather this valuable record. It is a massive collection of metadata and links between scientific products such as articles, datasets, software, and other research products, entities like organisations, funders, funding streams, projects, communities, and data sources.

As of today, the OpenAIRE Graph aggregates around 450Mi metadata records with links collecting from 2K data sources trusted by scientists, including:

- Open Access journals registered in DOAJ
- Crossref
- Unpaywall
- ORCID
- Microsoft Academic Graph
- Datacite

And repositories registered in OpenDOAR, re3data.org, FAIRSharing.org, and the EOSC Service Catalogue. Among these, prominent repositories such as:

- UKPubMed
- ArXiv
- HAL
- Zenodo
- Figshare
- Dryad
- Repec

After cleaning, deduplication, enrichment and full-text mining processes, the graph is analysed to produce statistics for the OpenAIRE MONITOR, the Open Science Observatory, made discoverable via the OpenAIRE EXPLORE and programmatically accessible via OpenAIRE Public APIs. Last but not least, the Graph data are openly available and can be used by third-parties to create added value services.

This technical Report describes the public data model adopted by the OpenAIRE Graph.

Data model

The OpenAIRE Graph comprises several types of entities and relationships among them.

The latest version of the JSON schema can be found on the Downloads section.

Entities

The main entities of the OpenAIRE Graph are listed in the following paragraph.

Result

Results are intended as digital objects, described by metadata, resulting from a scientific process. In this page, we describe the properties of the `Result` object.

Moreover, there are the following sub-types of a `Result`, that inherit all its properties and further extend it:

- Publication
- Dataset
- Software
- Other Research Product

id

Type: String • Cardinality: ONE

Main entity identifier, created according to the [OpenAIRE entity identifier and PID mapping policy]().

```
"id": "50|doi_dedup___::80f29c8c8ba18c46c88a285b7e739dc3"
```

type

Type: String • Cardinality: ONE

Type of the result. Possible types:

- `publication`
- `dataset`
- `software`
- `other`

as declared in the terms from the [dnet:result_typologies vocabulary]().

```
"type": "publication"
```

originalId

Type: String • Cardinality: MANY

Identifiers of the record at the original sources.

```
"originalId": [
    "oai:pubmedcentral.nih.gov:8024784",
    "S0048733321000305",
    "10.1016/j.respol.2021.104226",
    "3136742816"
]
```

maintitle

Type: String • Cardinality: ONE

A name or title by which a scientific result is known. May be the title of a publication, of a dataset or the name of a piece of software.

```
"maintitle": "The fall of the innovation empire and its possible rise thro
ugh open science"
```

subtitle

Type: String • Cardinality: ONE

Explanatory or alternative name by which a scientific result is known.

```
"subtitle": "An analysis of cases from 1980 - 2020"
```

author

Type: Author • Cardinality: MANY

The main researchers involved in producing the data, or the authors of the publication.

```
"author": [
    {
        "fullname": "E. Richard Gold",
        "rank": 1,
        "name": "Richard",
        "surname": "Gold",
        "pid": {
            "id": {
                "scheme": "orcid",
                "value": "0000-0002-3789-9238"
            },
            "provenance"; {
                "provenance": "Harvested",
                "trust": "0.9"
            }
        }
    },
    ...
]
```

bestaccessright

Type: BestAccessRight • Cardinality: ONE

The most open access right associated to the manifestations of this research results.

```
"bestaccessright": {
    "code": "c_abf2",
    "label": "OPEN",
    "scheme": "http://vocabularies.coar-repositories.org/documentation/acc
ess_rights/"
}
```

contributor

Type: String • Cardinality: MANY

The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.

```
"contributor": [
    "University of Zurich",
    "Wright, Aidan G C",
    "Hallquist, Michael",
    ...
]
```

country

Type: ResultCountry • Cardinality: MANY

Country associated with the result because it is the country of the organisation that manages the institutional repository or national aggregator or CRIS system from which this record was collected Country of affiliations of authors can be found instead in the affiliation rel.

```
"country": [
    {
        "code": "CH",
        "label": "Switzerland",
        "provenance": {
            "provenance": "Inferred by OpenAIRE",
            "trust": "0.85"
        }
    },
    ...
]
```

coverage

Type: String • Cardinality: MANY

dateofcollection

Type: String • Cardinality: ONE

When OpenAIRE collected the record the last time.

```
"dateofcollection": "2021-06-09T11:37:56.248Z"
```

description

Type: String • Cardinality: MANY

A brief description of the resource and the context in which the resource was created.

```
"description": [
    "Open science partnerships (OSPs) are one mechanism to reverse declini
ng efficiency. OSPs are public-private partnerships that openly share publ
ications, data and materials.",
    "There is growing concern that the innovation system's ability to crea
te wealth and attain social benefit is declining in effectiveness. This ar
ticle explores the reasons for this decline and suggests a structure, the
open science partnership, as one mechanism through which to slow down or r
everse this decline.",
    "The article examines the empirical literature of the last century to
document the decline. This literature suggests that the cost of research a
nd innovation is increasing exponentially, that researcher productivity is
declining, and, third, that these two phenomena have led to an overall fla
t or declining level of innovation productivity.",
    ...
]
```

embargoenddate

Type: String • Cardinality: ONE

Date when the embargo ends and this result turns Open Access.

```
"embargoenddate": "2017-01-01"
```

indicators

Type: Indicator • Cardinality: ONE

The indicators computed for this result; currently, the following two types of indicators are supported: impact indicators and usage statistics indicators.

```
"indicators": {
        "impactMeasures": {
                "influence": {
                        "score": "123",
                        "class": "C2"
                },
                "influence_alt" : {
                        "score": "456",
                        "class": "C3"
                },
                "popularity": {
                        "score": "234",
                        "class": "C1"
                },
                "popularity_alt": {
                        "score": "345",
                        "class": "C5"
                },
                "impulse": {
                        "score": "987",
                        "class": "C3"
                }
        },
        "usageCounts": {
                "downloads": "10",
                 "views": "20"
        }
}
```

Instance

Type: Instance • Cardinality: MANY

Specific materialization or version of the result. For example, you can have one result with three instances: one is the pre-print, one is the post-print, one is the published version.

```
"instance": [
    {
        "accessright": {
            "code": "c_abf2",
            "label": "OPEN",
            "openAccessRoute": "gold",
            "scheme": "http://vocabularies.coar-repositories.org/documenta
tion/access_rights/"
        },
        "alternateIdentifier": [
            {
                "scheme": "doi",
                "value": "10.1016/j.respol.2021.104226"
            },
            ...
        ],
        "articleprocessingcharge": {
            "amount": "4063.93",
            "currency": "EUR"
        },
        "license": "http://creativecommons.org/licenses/by-nc/4.0",
        "pid": [
            {
                "scheme": "pmc",
                "value": "PMC8024784"
            },
            ...
        ],

        "publicationdate": "2021-01-01",
        "refereed": "UNKNOWN",
        "type": "Article",
        "url": [
            "http://europepmc.org/articles/PMC8024784"
        ]
    },
    ...
]
```

language

Type: Language • Cardinality: ONE

The alpha-3/ISO 639-2 code of the language. Values controlled by the dnet:languages vocabulary.

```
"language": {
    "code": "eng",
    "label": "English"
}
```

lastupdatetimestamp

Type: Long • Cardinality: ONE

Timestamp of last update of the record in OpenAIRE.

```
"lastupdatetimestamp": 1652722279987
```

pid

Type: ResultPid • Cardinality: MANY

Persistent identifiers of the result. See also the OpenAIRE entity identifier and PID mapping policy to learn more.

```
"pid": [
    {
        "scheme": "pmc",
        "value": "PMC8024784"
    },
    {
        "scheme": "doi",
        "value": "10.1016/j.respol.2021.104226"
    },
    ...
]
```

publicationdate

Type: String • Cardinality: ONE

Main date of the research product: typically the publication or issued date. In case of a research result with different versions with different dates, the date of the result is selected as the most frequent well-formatted date. If not available, then the most recent and complete date among those that are well-formatted. For statistics, the year is extracted and the result is counted only among the result of that year. Example: Pre-print date: 2019-02-03, Article date provided by repository: 2020-02, Article date provided by Crossref: 2020, OpenAIRE will set as date 2019-02-03, because it's the most recent among the complete and well-formed dates. If then the repository updates the metadata and set a complete date (e.g. 2020-02-12), then this will be the new date for the result because it becomes the most recent most complete date. However, if OpenAIRE then collects the pre-print from another repository with date 2019-02-03, then this will be the “winning date” because it becomes the most frequent well-formatted date.

```
"publicationdate": "2021-03-18"
```

publisher

Type: String • Cardinality: ONE

The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource.

```
"publisher": "Elsevier, North-Holland Pub. Co"
```

Source

Type: String • Cardinality: MANY

A related resource from which the described resource is derived. See definition of Dublin Core field dc:source.

```
"source": [
      "Research Policy",
      "Crossref",
      ...
]
```

Subjects

Type: Subject • Cardinality: MANY

Subject, keyword, classification code, or key phrase describing the resource.

```
"subjects": [
    {
        "provenance": {
            "provenance": "Harvested",
            "trust": "0.9"
        },
        "subject": {
            "scheme": "keyword",
            "value": "Open science"
        }
    },
    ...
]
```

Sub-types

There are the following sub-types of `Result`. Each inherits all its fields and extends them with the following.

Publication

Metadata records about research literature (includes types of publications listed here).

container

Type: Container • Cardinality: ONE

Container has information about the conference or journal where the result has been presented or published.

```
"container": {
    "edition": "",
    "iss": "5",
    "issnLinking": "",
    "issnOnline": "1873-7625",
    "issnPrinted": "0048-7333",
    "name": "Research Policy",
    "sp": "12",
    "ep": "22",
    "vol": "50"
}
```

Dataset

Metadata records about research data (includes the subtypes listed here).

size

Type: String • Cardinality: ONE

The declared size of the dataset.

```
"size": "10129818"
```

version

Type: String • Cardinality: ONE

The version of the dataset.

```
"version": "v1.3"
```

geolocation

Type: GeoLocation • Cardinality: MANY

The list of geolocations associated with the dataset.

```
"geolocation": [
    {
        "box": "18.569386 54.468973  18.066832 54.83707",
        "place": "Tübingen, Baden-Württemberg, Southern Germany",
        "point": "7.72486 50.1084"
    },
    ...
]
```

Software

Metadata records about research software (includes the subtypes listed here).

documentationUrl

Type: String • Cardinality: MANY

The URLs to the software documentation.

```
"documentationUrl": [
    "https://github.com/openaire/iis/blob/master/README.markdown",
    ...
]
```

codeRepositoryUrl

Type: String • Cardinality: ONE

The URL to the repository with the source code.

```
"codeRepositoryUrl": "https://github.com/openaire/iis"
```

programmingLanguage

Type: String • Cardinality: ONE

The programming language.

```
"programmingLanguage": "Java"
```

Other research product

Metadata records about research products that cannot be classified as research literature, data or software (includes types of products listed here).

contactperson

Type: String • Cardinality: MANY

Information on the person responsible for providing further information regarding the resource.

```
"contactperson": [
    "Noémie Dominguez",
    ...
]
```

contactgroup

Type: String • Cardinality: MANY

Information on the group responsible for providing further information regarding the resource.

```
"contactgroup": [
    "Networked Multimedia Information Systems (NeMIS)",
    ...
]
```

tool

Type: String • Cardinality: MANY

Information about tool useful for the interpretation and/or re-use of the research

Data source

OpenAIRE entity instances are created out of data collected from various data sources of different kinds, such as publication repositories, dataset archives, CRIS systems, funder databases, etc. Data sources export information packages (e.g., XML records, HTTP responses, RDF data, JSON) that may contain information on one or more of such entities and possibly relationships between them.

For example, a metadata record about a project carries information for the creation of a Project entity and its participants (as Organization entities). It is important, once each piece of information is extracted from such packages and inserted into the OpenAIRE information space as an entity, for such pieces to keep provenance information relative to the originating data source. This is to give visibility to the data source, but also to enable the reconstruction of the very same piece of information if problems arise.

The `DataSource` object

id

Type: String • Cardinality: ONE

The OpenAIRE id of the data source, created according to the [OpenAIRE entity identifier and PID mapping policy](#).

```
"id": "10|issn___print::22c514d022b199c346e7f29ca06efc95"
```

originalId

Type: String • Cardinality: MANY

The list of original identifiers associated to the datasource.

```
"originalId": [
    "issn___print::2451-8271",
    ...
]
```

pid

Type: ControlledField • Cardinality: MANY

The persistent identifiers for the datasource.

```
"pid": [
    {
        "scheme": "DOI",
        "value": "10.5281/zenodo.4707307"
    },
    ...
]
```

datasourcetype

Type: ControlledField • Cardinality: ONE

The datasource type; see the vocabulary dnet:datasource_typologies.

```
"datasourcetype": {
    "scheme": "pubsrepository::journal",
    "value": "Journal"
}
```

openairecompatibility

Type: String • Cardinality: ONE

The OpenAIRE compatibility of the ingested results, indicates which guidelines they are compliant according to the vocabulary dnet:datasourceCompatibilityLevel.

```
"openairecompatibility": "collected from a compatible aggregator"
```

officialname

Type: String • Cardinality: ONE

The official name of the datasource.

```
"officialname": "Recent Patents and Topics on Medical Imaging"
```

englishname

Type: String • Cardinality: ONE

The English name of the datasource.

```
"englishname": "Recent Patents and Topics on Medical Imaging"
```

websiteurl

Type: String • Cardinality: ONE

The URL of the website of the datasource.

```
"websiteurl": "http://dspace.unict.it/"
```

logourl

Type: String • Cardinality: ONE

The URL of the logo for the datasource.

```
"logourl": "https://impactum-journals.uc.pt/public/journals/26/pageHeaderLogoImage_en_US.png"
```

dateofvalidation

Type: String • Cardinality: ONE

The date of validation against the OpenAIRE guidelines for the datasource records.

```
"dateofvalidation": "2016-10-10"
```

description

Type: String • Cardinality: ONE

The description for the datasource.

```
"description": "Recent Patents on Medical Imaging publishes review and research articles, and guest edited single-topic issues on recent patents in the field of medical imaging. It provides an important and reliable source of current information on developments in the field. The journal is essential reading for all researchers involved in Medical Imaging."
```

subjects

Type: String • Cardinality: MANY

List of subjects associated to the datasource

```
"subjects": [
    "Medicine",
    "Imaging",
    ...
]
```

languages

Type: String • Cardinality: MANY

The languages present in the data source's content, as defined by OpenDOAR.

```
"languages":[
    "eng",
    ...
]
```

contenttypes

Type: String • Cardinality: MANY

Types of content in the data source, as defined by OpenDOAR

```
"contenttypes": [
    "Journal articles",
    ...
]
```

releasestartdate

Type: String • Cardinality: ONE

Releasing date of the data source, as defined by re3data.org.

```
"releasestartdate": "2010-07-24"
```

releaseenddate

Type: String • Cardinality: ONE

Date when the data source went offline or stopped ingesting new research data. As defined by re3data.org

```
"releaseenddate": "2016-03-28"
```

accessrights

Type: String • Cardinality: ONE

Type of access to the data source, as defined by re3data.org. Possible values: { `open, restricted, closed` }.

```
"accessrights": "open"
```

uploadrights

Type: String • Cardinality: ONE

Type of data upload, as defined by re3data.org; one of { `open, restricted, closed` }.

```
"uploadrights": "closed"
```

databaseaccessrestriction

Type: String • Cardinality: ONE

Access restrictions to the research data repository. Allowed values are: { `feeRequired, registration, other` }.

This field only applies for re3data data source; see re3data schema specification for more details.

```
"databaseaccessrestriction": "registration"
```

datauploadrestriction

Type: String • Cardinality: ONE

Upload restrictions applied by the datasource, as defined by re3data.org. One of { `feeRequired, registration, other` }.

This field only applies for re3data data source; see re3data schema specification for more details.

```
"datauploadrestriction": "feeRequired registration"
```

versioning

Type: Boolean • Cardinality: ONE

Whether the research data repository supports versioning: yes if the data source supports versioning, no otherwise.

This field only applies for re3data data source; see re3data schema specification for more details.

```
"versioning": true
```

citationguidelineurl

Type: String • Cardinality: ONE

The URL of the data source providing information on how to cite its items. The DataCite citation format is recommended (http://www.datacite.org/whycitedata).

This field only applies for re3data data source; see re3data schema specification for more details.

```
"citationguidelineurl": "https://physionet.org/about/#citation"
```

pidsystems

Type: String • Cardinality: ONE

The persistent identifier system that is used by the data source. As defined by re3data.org.

```
"pidsystems": "hdl"
```

certificates

Type: String • Cardinality: ONE

The certificate, seal or standard the data source complies with. As defined by re3data.org.

```
"certificates": "WDS"
```

policies

Type: String • Cardinality: MANY

Policies of the data source, as defined in OpenDOAR.

journal

Type: Container • Cardinality: ONE

Information about the journal, if this data source is of type Journal.

```
"container": {
    "edition": "",
    "iss": "5",
    "issnLinking": "",
    "issnOnline": "1873-7625",
    "issnPrinted":"2451-8271",
    "name": "Recent Patents and Topics on Imaging",
    "sp": "12",
    "ep": "22",
    "vol": "50"
}
```

missionstatementurl

Type: String • Cardinality: ONE

The URL of a mission statement describing the designated community of the data source. As defined by re3data.org

```
"missionstatementurl": "https://www.sigma2.no/content/nird-research-data-archive"
```

Organization

Organizations include companies, research centers or institutions involved as project partners or as responsible of operating data sources. Information about organizations are collected from funder databases like CORDA, registries of data sources like OpenDOAR and re3Data, and CRIS systems, as being related to projects or data sources.

The `Organization` object

id

Type: String • Cardinality: ONE

The OpenAIRE id for the organization, created according to the OpenAIRE entity identifier and PID mapping policy.

```
"id": "20|openorgs____::b84450f9864182c67b8611b5593f4250"
```

legalshortname

Type: String • Cardinality: ONE

The legal name in short form of the organization.

```
"legalshortname": "ARC"
```

legalname

Type: String • Cardinality: ONE

The legal name of the organization.

```
"legalname": "Athena Research and Innovation Center In Information Communi
cation & Knowledge Technologies"
```

alternativenames

Type: String • Cardinality: MANY

Alternative names that identify the organization.

```
"alternativenames": [
    "Athena Research and Innovation Center In Information Communication &
Knowledge Technologies",
    "Athena RIC",
    "ARC",
    ...
]
```

Project

Of crucial interest to OpenAIRE is also the identification of the funders (e.g. European Commission, WellcomeTrust, FCT Portugal, NWO The Netherlands) that co-funded the projects that have led to a given result. Projects are characterized by a list of funding streams (e.g. FP7, H2020 for the EC), which identify the strands of fundings. Funding streams can be nested to form a tree of sub-funding streams.

The `Project` object

id

Type: String • Cardinality: ONE

Main entity identifier, created according to the OpenAIRE entity identifier and PID mapping policy.

```
"id": "40|corda__h2020::70ea22400fd890c5033cb31642c4ae68"
```

code

Type: String • Cardinality: ONE

The grant agreement code of the project.

```
"code": "777541"
```

acronym

Type: String • Cardinality: ONE

Project’s acronym.

```
"acronym": "OpenAIRE-Advance"
```

title

Type: String • Cardinality: ONE

Project’s title.

```
"title": "OpenAIRE Advancing Open Scholarship"
```

callidentifier

Type: String • Cardinality: ONE

The identifier of the research call.

```
"callidentifier": "H2020-EINFRA-2017"`
```

funding

Type: Funding • Cardinality: MANY

Funding information for the project.

```
"funding": [
    {
        "funding_stream": {
            "description": "Horizon 2020 Framework Programme - Research an
d Innovation action",
            "id": "EC::H2020::RIA"
        },
        "jurisdiction": "EU",
        "name": "European Commission",
        "shortName": "EC"
    }
]
```

granted

Type: Grant • Cardinality: ONE

The money granted to the project.

```
"granted": {
    "currency": "EUR",
    "fundedamount": 1.0E7,
    "totalcost": 1.0E7
}
```

h2020programme

Type: H2020Programme • Cardinality: MANY

The H2020 programme funding the project.

```
"h2020programme":[
    {
        "code": "H2020-EU.1.4.1.3.",
        "description": "Development, deployment and operation of ICT-based
e-infrastructures"
    }
]
```

keywords

Type: String • Cardinality: ONE

```
"keywords": [
    "Open Science",
    ...
]
```

openaccessmandatefordataset

Type: Boolean • Cardinality: ONE

```
"openaccessmandatefordataset": true
```

openaccessmandateforpublications

Type: Boolean • Cardinality: ONE

```
"openaccessmandateforpublications": true
```

startdate

Type: String • Cardinality: ONE

The start year of the project.

```
"startdate": "2018-01-01"
```

enddate

Type: String • Cardinality: ONE

The end year pf the project.

```
"enddate": "2021-02-28"
```

subject

Type: String • Cardinality: MANY

The subjects of the project

```
"subject": [
    "Data and Distributed Computing e-infrastructures for Open Science",
    ...
]
```

summary

Type: String • Cardinality: ONE

Short summary of the project.

```
"summary": "OpenAIRE-Advance continues the mission of OpenAIRE to support 
the Open Access/Open Data mandates in Europe. By sustaining the current su
ccessful infrastructure, comprised of a human network and robust technical 
services, it consolidates its achievements while working to shift the mome
ntum among its communities to Open Science, aiming to be a trusted e-Infra
structurewithin the realms of the European Open Science Cloud.In this next 
phase, OpenAIRE-Advance strives to empower its National Open Access Desks 
(NOADs) so they become a pivotal part within their own national data infra
structures, positioningOA and open science onto national agendas. The capa
city building activities bring together experts ontopical task groups in t
hematic areas(open policies, RDM, legal issues, TDM), promoting a train th
e trainer approach, strengthening and expanding the pan-European Helpdesk 
with support and training toolkits, training resources and workshops.It ex
amines key elements of scholarly communication, i.e., co-operative OA publ
ishing and next generation repositories, to develop essential building blo
cks of the scholarly commons.On the technical level OpenAIRE-Advance focus
es on the operation and maintenance of the OpenAIRE technical TRL8/9 servi
ces,and radically improvesthe OpenAIRE services on offer by: a) optimizing 
their performance and scalability, b) refining their functionality based o
n end-user feedback, c) repackagingthem into products, taking a profession
al marketing approach  with well-defined KPIs, d)consolidating the range o
f services/products into a common e-Infra catalogue to enable a wider upta
ke.OpenAIRE-Advancesteps up its outreach activities with concrete pilots w
ith three major RIs,citizen science initiatives, and innovators via a rigo
rous Open Innovation programme. Finally, viaits partnership with COAR, Ope
nAIRE-Advance consolidatesOpenAIRE’s global roleextending its collaboratio
ns with Latin America, US, Japan, Canada, and Africa."
```

websiteurl

Type: String • Cardinality: ONE

The website of the project

```
"websiteurl": "https://www.openaire.eu/advance/"
```

Community

Research communities and research initiatives are intended as groups of people with a common research intent and can be of two types: research initiatives or research communities:

- Research initiatives are intended to capture a view of the information space that is “research impact”-oriented, i.e. all products generated due to my research initiative;
- Research communities the latter “research activity” oriented, i.e. all products that may be of interest or related to my research initiative.

For example, the organizations supporting a research infrastructure fall in the first category, while the researchers involved in a discipline fall in the second.

The `Community` object

id

Type: String • Cardinality: ONE

The OpenAIRE id for the community/research infrastructure, created according to the OpenAIRE entity identifier and PID mapping policy.

```
"id": "00|context_____::5b7f9fa40bdc12072249204cedfa7808"
```

acronym

Type: String • Cardinality: ONE

The acronym of the community.

```
"acronym": "covid-19"
```

description

Type: String • Cardinality: ONE

Description of the research community/research infrastructure

```
"description": "This portal provides access to publications, research data
, projects and software that may be relevant to the Corona Virus Disease (
COVID-19). The OpenAIRE COVID-19 Gateway aggregates COVID-19 related recor
ds, links them and provides a single access point for discovery and naviga
tion. We tag content from the OpenAIRE Graph (10,000+ data sources) and ad
ditional sources. All COVID-19 related research results are linked to peop
le, organizations and projects, providing a contextualized navigation."
```

name

Type: String • Cardinality: ONE

The long name of the community.

```
"name": "Corona Virus Disease"
```

subject

Type: String • Cardinality: MANY

The list of the subjects associated to the research community (only appies to research communities).

```
"subject": [
    "COVID19",
    "SARS-CoV",
    "HCoV-19",
    ...
]
```

type

Type: String • Cardinality: ONE

The type of the community; one of `{ Research Community, Research infrastructure }`.

```
"type": "Research Community"
```

zenodo_community

Type: String • Cardinality: ONE

The URL of the Zenodo community associated to the Research community/Research infrastructure.

```
"zenodo_community": "https://zenodo.org/communities/covid-19"
```

Relationships

A relationship in the graph is represented by the following data type, which aims to model a directed edge between two nodes, providing information about the semantic of the relation, its provenance and validation.

The `Relationship` object

source

Type: Node • Cardinality: ONE

Represents the source node in the relation.

```
"source": {
    "id": "20|openorgs____::1cb75a3ad756e4c83e455e3e7347643b",
    "type": "organization"
}
```

target

Type: Node • Cardinality: ONE

Represents the target node in the relation.

```
"target": {
    "id": "10|doajarticles::022409068174087a003647ff46070f7f",
    "type": "datasource"
}
```

reltype

Type: RelType • Cardinality: ONE

Represent the semantics of the relation between two nodes of the graph.

```
"reltype": {
    "name": "provides",
    "type": "provision"
}
```

provenance

Type: Provenance • Cardinality: ONE

Indicates the process that produced (or provided) the information.

```
"provenance": {
    "provenance": "Harvested",
    "trust":"0.900"
}
```

validated

Type: Boolean • Cardinality: ONE

Indicates weather or not the relation was validated.

```
"validated": true
```

validationDate

Type: String • Cardinality: ONE

Indicates the validation date of the relation - applies only when the validated flag is set to true.

```
"validationDate": "2022-09-02"
```

The `Node` object

The Node data type contains the minimum information needed to identify a graph node, its identifier and entity type.

id

Type: String • Cardinality: ONE

OpenAIRE identifier of the node in the graph.

```
"id": "10|doajarticles::022409068174087a003647ff46070f7f"
```

type

Type: String • Cardinality: ONE

Graph node type.

```
"type": "datasource"
```

The `RelType` object

The RelType data type models the semantic of the relationship among two nodes.

type

Type: String • Cardinality: ONE

Relation category, e.g. affiliation, citation, see table Relation typologies.

```
"name": "provides"
```

name

Type: String • Cardinality: ONE

Further specifies the relation semantic, indicating the relation direction, e.g. Cites, isCitedBy.

```
"type": "provision"
```

Relationship types

The following table lists all the possible relation semantics found in the graph dump.

Note: the labels used to specify the semantic of the relationships are (for the large) inherited from the DataCite metadata kernel, which provides a description for them.

#	Source entity type	Target entity type	Relation name / inverse	Provenance
1	Project	Result	produces / isProducedBy	Harvested, Inferred by OpenAIRE, Linked by user
2	Project	Organization	hasParticipant / isParticipant	Harvested
3	Project	Community	IsRelatedTo / IsRelatedTo	Linked by user
4	Result	Result	IsAmongTopNSimilarDocuments / HasAmongTopNSimilarDocuments	Inferred by OpenAIRE
5	Result	Result	IsSupplementTo / IsSupplementedBy	Harvested
6	Result	Result	IsRelatedTo / IsRelatedTo	Harvested, Inferred by OpenAIRE, Linked by user
7	Result	Result	IsPartOf / HasPart	Harvested
8	Result	Result	IsDocumentedBy / Documents	Harvested
9	Result	Result	IsObsoletedBy / Obsoletes	Harvested
10	Result	Result	IsSourceOf / IsDerivedFrom	Harvested
11	Result	Result	IsCompiledBy / Compiles	Harvested
12	Result	Result	IsRequiredBy / Requires	Harvested
13	Result	Result	IsCitedBy / Cites	Harvested, Inferred by OpenAIRE
14	Result	Result	IsReferencedBy / References	Harvested
15	Result	Result	IsReviewedBy / Reviews	Harvested

16	Result	Result	IsOriginalFormOf / IsVariantFormOf	Harvested
17	Result	Result	IsVersionOf / HasVersion	Harvested
18	Result	Result	IsIdenticalTo / IsIdenticalTo	Harvested
19	Result	Result	IsPreviousVersionOf / IsNewVersionOf	Harvested
20	Result	Result	IsContinuedBy / Continues	Harvested
21	Result	Result	IsDescribedBy / Describes	Harvested
22	Result	Organization	hasAuthorInstitution / isAuthorInstitutionOf	Harvested, Inferred by OpenAIRE
23	Result	Data source	isHostedBy / hosts	Harvested, Inferred by OpenAIRE
24	Result	Data source	isProvidedBy / provides	Harvested
25	Result	Community	IsRelatedTo / IsRelatedTo	Harvested, Inferred by OpenAIRE, Linked by user
26	Organization	Community	IsRelatedTo / IsRelatedTo	Linked by user
27	Organization	Organization	IsChildOf / IsParentOf	Linked by user
28	Data source	Community	IsRelatedTo / IsRelatedTo	Linked by user
29	Data source	Organization	isProvidedBy / provides	Harvested

PIDs and identifiers

One of the challenges towards the stability of the contents in the OpenAIRE Graph consists of making its identifiers and records stable over time. The barriers to this scenario are many, as the Graph keeps a map of data sources that is subject to constant variations: records in repositories vary in content, original IDs, and PIDs, may disappear or reappear, and the same holds for the repository or the metadata collection it exposes. Not only, but the mappings applied to the original contents may also change and improve over time to catch up with the changes in the input records.

PID Authorities

One of the fronts regards the attribution of the identity to the objects populating the graph. The basic idea is to build the identifiers of the objects in the graph from the PIDs available in some authoritative sources while considering all the other sources as by definition “unstable”. Examples of authoritative sources are Crossref and DataCite. Examples of non-authoritative ones are institutional repositories, aggregators, etc. PIDs from the authoritative sources would form the stable OpenAIRE ID skeleton of the Graph, precisely because they are immutable by construction.

Such a policy defines a list of data sources that are considered authoritative for a specific type of PID they provide, whose effect is twofold: * OpenAIRE IDs depend on persistent IDs when they are provided by the authority responsible to create them; * PIDs are included in the graph according to a tight criterion: the PID Types declared in the table below are considered to be mapped as PIDs only when they are collected from the relative PID authority data source.

PID Type	Authority
doi	Crossref, Datacite
pmc, pmid	Europe PubMed Central, PubMed Central
arXiv	arXiv.org e-Print Archive

There is an exception though: Handle(s) are minted by several repositories; as listing them all would not be a viable option, to avoid losing them as PIDs, Handles bypass the PID authority filtering rule. In all other cases, PIDs are be included in the graph as alternate Identifiers.

Delegated authorities

When a record is aggregated from multiple sources considered authoritative for minting specific PIDs, different mappings could be applied to them and, depending on the case, this could result in inconsistencies in the attribution of the field values. To overcome the issue, the intuition is to include such records only once in the graph. To do so, the concept of "delegated authorities" defines a list of datasources that assigns PIDs to their scientific products from a given PID minter.

This "selection" can be performed when the entities in the graph sharing the same identifier are grouped together. The list of the delegated authorities currently includes

Datasource delegated	Datasource delegating	Pid Type
Zenodo	Datacite	doi
RoHub	W3ID	w3id

Identifiers in the Graph

OpenAIRE assigns internal identifiers for each object it collects. By default, the internal identifier is generated as `sourcePrefix::md5(localId)` where:

- `sourcePrefix` is a namespace prefix of 12 chars assigned to the data source at registration time
- `localid` is the identifier assigned to the object by the data source

After years of operation, we can say that:

- `localId` are generally unstable
- objects can disappear from sources
- PIDs provided by sources that are not PID agencies (authoritative sources for a specific type of PID) are often wrong (e.g. pre-print with the DOI of the published version, DOIs with typos)

Therefore, when the record is collected from an authoritative source:

- the identity of the record is forged using the PID, like `pidTypePrefix::md5(lowercase(doi))`
- the PID is added in a `pid` element of the data model

When the record is collected from a source which is not authoritative for any type of PID: * the identity of the record is forged as usual using the local identifier * the PID, if available, is added as `alternateIdentifier`

Currently, the following data sources are used as “PID authorities”:

PID Type	Prefix (12 chars)	Authority
doi	`doi_________`	Crossref, Datacite, Zenodo
pmc	`pmc_________`	Europe PubMed Central, PubMed Central
pmid	`pmid________`	Europe PubMed Central, PubMed Central
arXiv	`arXiv_______`	arXiv.org e-Print Archive
handle	`handle______`	any repository

OpenAIRE also perform duplicate identification (see the dedicated section for details). All duplicates are **merged** together in a **representative record** which must be assigned a dedicated OpenAIRE identifier (i.e. it cannot have the identifier of one of the aggregated record).

Table of Contents