


Case Study: Principles for high-quality data sources in OpenAIRE

Summary	Encoding of the principles and criteria for OPENAIRE data source quality in CAT.
Audience	
Status	Public Draft

Date	Authors	Description	Version
24-02-2024	W Hugo	Inception and first draft	0.1
31-05-2024	P Manghi	Elaboration of OpenAIRE principles, criteria, tests, and metrics	0.2
28-05-2025	W Hugo	Editorial corrections and publication	1.0

Licence	CC 4.0 BY		
Funding	European Union Funded Project 101057264 — FAIRCORE4EOSC	 Funded by the European Union	
Contributors	W Hugo (DANS), P Manghi (OpenAIRE)		

Context

Data sources are services operated to facilitate sharing and access of scholarly outputs, such as publications, research data and research software. The OpenAIRE Graph is a free and open resource that brings together and interlinks more than 250Mi of scholarly output metadata records from over 100k data sources trusted by researchers. To be integrated and aggregated into the OpenAIRE Graph, data sources must comply to the OpenAIRE Guidelines, hence respect the interoperability formats and protocols specified therein. This harmonization allows for efficient data exchange between repositories, funding agencies, and other scholarly communication platforms, fostering a more connected and collaborative research environment.

In the rapidly evolving landscape of scholarly communication, the adherence to the [OpenAIRE Guidelines](#) by data sources holds paramount importance. By onboarding metadata to the OpenAIRE Graph, data sources contribute to a global, interconnected collection of scientific knowledge that accelerates discovery, innovation, interoperability, accessibility, and monitoring of science across various disciplines and geographical boundaries.

FAIRness Compliance with the guidelines enhances the visibility, accessibility, and reproducibility of research outputs. Metadata standards prescribed by OpenAIRE ensure that research data is easily discoverable by search engines and indexing services. This increased discoverability means that researchers, policymakers, and the public can easily find and access relevant research, thereby amplifying the reach and impact of scholarly work.

Adoption of standards OpenAIRE Guidelines promote the use of standardized metadata formats and controlled vocabularies, ensuring the quality and consistency of data. High-quality, standardized data is crucial for reliable search, retrieval, and analysis, contributing to the integrity and reproducibility of research.

Funder mandates The EC requires that the outputs of funded projects are deposited in OpenAIRE compliant repositories to ensure Open Access, project reporting, and monitoring, as well as to be onboarded into the Resource Hub of the EOSC EC Node. Data sources that follow OpenAIRE Guidelines help researchers comply with these mandates, simplifying the process of grant reporting and funder compliance.

Adhering to the OpenAIRE Guidelines is not merely a technical requirement but also a strategic choice that amplifies the impact of research outputs and fosters a more open, accessible, and collaborative scholarly communication infrastructure. By ensuring interoperability, visibility, and quality, data sources can significantly enhance the dissemination and utilization of scholarly knowledge, ultimately driving forward scientific progress and societal advancement.

This document describes the actors and the policy engaged in the integration of data sources into the OpenAIRE Graph. To this aim, it provides a specification according to the CAT service as implemented in the FAIRCORE4EOSC project. The attempt is to validate the framework while providing a formal specification of the criteria and tests that actors should meet to make sure their data source complies to the [OpenAIRE Guidelines](#).

Actors

- **Provider:** The organization that is liable for and ensures the operation of a Data Source.
- **Data Source:** A Service that supports deposition, storage, and access to metadata and files relative to digital research products, such as publications, datasets, and software. A Data Source can be operated by one or more **Providers**
- **OpenAIRE Graph:** An SKG Service operated by OpenAIRE, capable of aggregating metadata records from **Data Sources** to populate an Open Science Knowledge Graph.

Principles/ Objectives

To define a policy a set of principles must be expressed (the intended beneficial outcomes of the policy).

#	Principle or Objective	Description
P-1	Accessibility	Registration to registry: the Provider should make sure the Data Source information is made available for discovery and access to potential consumers
P-2	Interoperability	The Provider that is willing to aggregate its Data Source into the OpenAIRE Graph must ensure the Data Source conforms to the OpenAIRE Interoperability Guidelines
P-3	Sustainability	The Provider must guarantee and provide evidence of the sustainability plan of the Data Source
P-4	Reliability	The Provider must ensure a high Quality of Service, to ensure regular aggregation of metadata and files from the OpenAIRE Graph

Criteria/Provisions

Each principle/objective is implemented by a criterion.

Should show alignment with or fulfilment of principles/objectives.

- a. Which actor(s) needs to align or confirm;
- b. Is it mandatory or optional (and maybe conditional)
- c. Should be clear how conformance will be determined

For each criterion a metric should be defined, that metric should match a benchmark. Maturity levels would be an example of comparing a performance level (metric value) with a benchmark.

#	Principle or Objective	Suggested criterion	Description	Metric	Benchmark
C1.1	P-1 Accessibility	Registration to registry	Providers must ensure the Data Source is described by a profile in a repository registry among FAIRSharing.org, OpenDOAR.org, and re3data.org	T1	=0 → 0 =1 → 1
C2.1	P-2 Interoperability	Registration to OpenAIRE	Providers must register the Data Source to the OpenAIRE Graph, providing the OAI-PMH APIs if these are not available (or incorrect) from the Data Source profile	T2	=0 → 0 =1 → 1
C2.2	P-2 Interoperability	Validation of basic OAI-PMH compliance	Providers must perform and pass a validation test to verify compliance of the Data Source to OAI-PMH refresh harvesting mode	T3	=0 → 0 =1 → 1
C2.3	P-2 Interoperability	Validation of full OAI-PMH compliance	Providers must perform and pass a validation test to verify compliance of the Data Source to OAI-PMH full protocol	T4	=0 → 0 =1 → 1
C2.4	P-2 Interoperability	Validation of compliance to OpenAIRE Guidelines 4.0 metadata schema	Providers must perform and pass a validation test to verify compliance of metadata records to the OpenAIRE Guidelines, in terms of structure (fields) and semantics	T5	<a → 0 =>a → 1

			(vocabularies and formats)		
C2.5	P-2 Interoperability	Availability to share the files described by the metadata	Providers can allow downloading of files described by metadata records	T6	=0 → 0 =1 → 1
C3.1	P-3 Sustainability	Sustainability of the Data Source	The Data Source has a long-term operational plan	T7 or T9	=0 → 0 =1 → 1
C4.1	P-4 Reliability	Reliability of Data Source APIs	The Data Source API can stand the aggregation requests ratio from the OpenAIRE Graph aggregation system	T8 or T10	=0 → 0 =1 → 1

Tests

One or more tests can compose a metric. Tests are reusable for more than one metric in some cases. Tests can be simple (e.g. yes-no) or more complex.

#	Test	Description	Type	Method	Guidance
T1	PROFILE	Login to OpenAIRE PROVIDE service and attempt registration. If the Data Source profile is not made available by the service, then the Data Source is not registered to FAIRSharing, OpenDOAR, or re3data profiles	Binary	Yes = 1 No = 0	G1
T2	REGISTRATION	Login to OpenAIRE PROVIDE and verify that the Data Source has an entry in the service	Binary	Yes = 1 No = 0	G1
T3	OAI_PMH_basic	Login to OpenAIRE PROVIDE and test OAI-PMH interface for the Data Source, verify OAI-PMH refresh option passes the test	Binary	Yes = 1 No = 0	G2
T4	OAI_PMH_full	Login to OpenAIRE PROVIDE and test OAI-PMH interface for the Data Source, verify all protocol verbs pass the test	Binary	Yes = 1 No = 0	G2
T5	METADATA_4.0	Login to OpenAIRE PROVIDE and validate metadata against the OpenAIRE Guidelines 4.0	Value	<a → 0 =>a → 1	G3
T6	FILE_DOWNLOAD	Verify Provider has agreed to file download during registration of Data Source	Binary	Yes = 1 No = 0	G4

T7	SUSTAINABILITY_PROACTIVE	Find Data Source information that proves sustainability	Binary	Yes = 1 No = 0	G5
T8	RELIABILITY_PROACTIVE	Find Data Source accounting data that proves reliability	Binary	Yes = 1 No = 0	G6
T9	SUSTAINABILITY_DECLARATION	Request Provider declaration of sustainability intention	Binary	Yes = 1 No = 0	G7
T10	RELIABILITY_DECLARATION	Request Provider declaration of reliability insurance	Binary	Yes = 1 No = 0	G8

Guidance

#	Guidance
G1	Access OpenAIRE PROVIDE (http://provide.openaire.eu) and use the Wizard for the registration of the Data Source to OpenAIRE
G2	Access OpenAIRE PROVIDE (http://provide.openaire.eu) and perform validation of OAI-PMH protocols and verify the protocol's score in the report
G3	Access OpenAIRE PROVIDE (http://provide.openaire.eu), perform validation of the Data Source against guidelines, and verify the score "a" of the report
G4	Inspect Data Source profiles, web site, and documentation to make sure the Provider has acceptable operation sustainability plans
G5	Find Data Source accounting data that proves sustainability, e.g. visit data source web site
G6	Find Data Source monitoring data that proves reliability, e.g. EOSC monitoring services of Data Source is integrated with EOSC Monitor
G7	Access OpenAIRE PROVIDE (http://provide.openaire.eu) and verify if the Provider has signed a Sustainability Commitment declaration
G8	Access OpenAIRE PROVIDE (http://provide.openaire.eu) and verify if the Provider has signed a Service Level Agreement

Applicability of Criteria to Actors

#	Criterion	Imperative	Provider	Data Source
C1.1	Registration to registry	MUST		✓
C2.1	Registration to OpenAIRE	MUST		✓
C2.2	Validation of basic OAI-PMH compliance	MUST		✓
C2.3	Validation of full OAI-PMH compliance	SHOULD		✓
C2.4	Validation of compliance to OpenAIRE Guidelines 4.0 metadata schema	MUST		✓
C2.5	Availability to share the files described by the metadata	MAY	✓	
C3.1	Sustainability of the Data Source	SHOULD	✓	
C4.1	Reliability of Data Source APIs	SHOULD		✓