

OpenAIRE Content Acquisition Policy: expanding the scope

Pedro Príncipe, University of Minho, pedroprincipe@sdum.uminho.pt

Session Type

- Poster

Abstract

This poster outlines the new OpenAIRE Content Acquisition Policy released in October 2018, which 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.

With its new content acquisition policy, OpenAIRE broadens its scope to integrate metadata records of all scientific and research products. This means that OpenAIRE now harvests: publication records of all access levels (open access, closed access/metadata only etc.), publication records with and without funding references, publication records of different research product types in one repository (literature publications, research data, software and other research products). In order to grant that records are included in OpenAIRE, it is vital that the access level of a record is made clear (preferably by an access level statement on record level, alternately by the use of specific OAI-sets) and each record contains a PID (or URL) that resolves to a splash page.

Conference Themes

- Open and sustainable
- Policies, licensing and the law
- Repositories and global knowledge

Keywords

OpenAIRE graph, content policy, repositories infrastructure

Audience

Repository managers, developers, librarians

Background

The new Content Acquisition Policy 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.

It is important to stress that OpenAIRE collects publications from EC-funded projects as well as every other Open Access (OA) content. In this new Content Acquisition Policy version, OpenAIRE is expanding the scope again to include non-OA content (under certain conditions) too. Specifically,

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). Please note that OpenAIRE do not exclude non-OA content of repositories compliant with former versions of the guidelines.

Presentation content

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 favor 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
- 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 and their import in OpenAIRE is curated by OpenAIRE data curators. Check out all kinds of data providers or just 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.

Conclusion

The OpenAIRE graph links information about millions of research outputs (such as publications, datasets, software etc.) from various entities (e.g. organizations and projects). It is based on an open set of metadata of interlinked scientific items. These can include literature, datasets, software. All of these items are linked to their open access status, funding information and research community status. The graph is made unique because of the following features:

1. Complete: collection of metadata about all scientific output, with corresponding links.
2. Participatory: relying on this infrastructure's growing network of content providers, repositories, publishers, infrastructures.
3. Cleaned and deduplicated: duplication is inevitable, but can be overcome by applying new mechanisms for de-duplication.

This key 'kernel' of information is structured around the new content acquisition policy which now covers all literature (both open and closed) in short covering 'all' of scholarship. The numbers have grown exponentially since December 2018, aiming for a coverage of 100 million publications and 10 million datasets.

References

Becker, Amelie, Loeden, Aenne, Manghi, Paolo, Principe, Pedro, & Schirrwagen, Jochen. (2018). OpenAIRE Content Acquisition Policies (Version 1.0).
<https://doi.org/10.5281/zenodo.1446407>