

# The OpenAIRE ScholeXplorer Service: aggregation and resolution of literature-dataset links

*Tuesday, 9 October 2018 14:30 (15 minutes)*

**OpenAIRE** OpenAIRE is the European infrastructure in support of Open Science. It fosters and monitors the adoption of Open Science across Europe and beyond, at the National and international level and at the research community level. It advocates the importance and the uptake of Open Science-oriented research life-cycles and publishing workflows, in support of reproducible science, transparent assessment, and omni-comprehensive scientific reward. To this aim OpenAIRE leverages the required cultural shift via a pervasive network of people in Europe (NOADs - National Open Access Desks) and beyond ("global alignment" via CORE), and facilitates the technological shift by providing technical services and interoperability guidelines.

**Scholix** Under the international forum of the Research Data Alliance and in collaboration with relevant stakeholders in the field, such as DataCite, CrossRef, World Data System, and Elsevier, OpenAIRE has participated to a Working Group for the definition of the Scholix framework (Scholarly Link eXchange). The goal of Scholix is to establish a high level interoperability framework for exchanging information about the links between scholarly literature and data. It aims to enable an open information ecosystem to understand systematically what data underpins literature and what literature references data. Scholix maintains an evolving set of Guidelines consisting of: (i) an information model (conceptual definition of what is a Scholix scholarly link), (ii) a link metadata schema (set of metadata fields representing a Scholix link), and (iii) a corresponding XML and JSON schema. Scholix is currently adopted as export format for links by DataCite and CrossRef via the CrossRef EventData service, by EuropePMC, and by OpenAIRE via the ScholeXplorer service.

**ScholeXplorer** ScholeXplorer is an OpenAIRE production service that since 2017 offers access to a unique collection of links between publications and datasets collected from publishers (EventData), data centres (DataCite), and institutional and thematic repositories (OpenAIRE). The collection is constantly populated and features 31Mi bi-directional links between 880.000 articles and 5.840.000 datasets from an overall 13.000 providers. The resulting graph of links can be accessed via the ScholeXplorer portal or via the APIs, which support third-party services at resolving publication/dataset PIDs to obtained related datasets or publications - content is also made available as a JSON dump via Zenodo.org. Since the beginning of 2018 the service has counted around 700 Million requests for PID resolution by third-party services (mainly Elsevier ScienceDirect) which have integrated ScholeXplorer in their workflows to show the list of datasets (publications) linked to their publications (datasets).

In this presentation we shall present the benefits of Scholix and the technical challenges underlying ScholeXplorer as a production service (i.e. aggregation, resolution, deduplication of link metadata) and the solutions adopted to achieve the quality of service as agreed on with data centers and publishers, which are today using the service as their main link-exchange channel.

## Type of abstract

Presentation

## Summary

We shall present the technical challenges underlying the realization and operation of the OpenAIRE ScholeXplorer Service, a service resulting from the RDA Scholix Working Group. ScholeXplorer offers access to a unique collection of links between publications and datasets continuously collected from publishers (EventData), data centres (DataCite), and institutional and thematic repositories (OpenAIRE). The collection currently features 31Mi bi-directional links between 880.000 articles and 5.840.000 datasets from an overall 13.000 providers. Since the beginning of 2018 the service has counted around 700 Million requests for PID resolution by third-party services (mainly Elsevier ScienceDirect).

**Primary authors:** MANGHI, Paolo (Istituto di Scienza e Tecnologie dell'Informazione - CNR); Mr LA BRUZZO,

Sandro (ISTI-CNR)

**Presenter:** MANGHI, Paolo (Istituto di Scienza e Tecnologie dell'Informazione - CNR)

**Session Classification:** Open Science

**Track Classification:** Area 1. Cross-Domain challenges / Data exchange across domains: researchers, technologist and policy makers perspectives