OpenOrgs: the OpenAIRE tool for bridging registries of research organizations

Claudio Atzori[1], Gina Pavone[1], Ivana Končić[2], Bojan Macan[2]
[1]Institute of Information Science and Technologies, Italy
[2]Ruđer Bošković Institute, Croatia

ABSTRACT

Building a connected open scholarly communication system requires unambiguously identifying research-related entities. It is not a simple task: for example, the same organization can have a range of names (eg. legal name, a shortened version, an abbreviation, in national language or in English), as well as varied metadata in other sources. Additionally, persistent IDs might not be helpful when several data sources (eg. ROR, ISNI, EC PIC) utilize various PID schemas to identify organizations. Due to this ambiguity, there are efficiency issues with information sharing, discoverability of research outputs, keeping track of activities, and ultimately building an integrated open scholarly communication system and OS services (Artini et al., 2022).

A new tool called OpenOrgs was developed to address this old issue: the disambiguation of organizations engaged in the research process (Pavone, 2021) as well as the parent-child relationships between departments and organizations. OpenOrgs tackles the ambiguity in the data that OpenAIRE collects from several research organization registries (eg. ROR), as well as other sources including institutional repositories, scientific journals, and CRIS systems, and aggregates them to populate the OpenAIRE Graph (OpenAIRE Graph, n.d.). To make up for the lack of information and increase the organization's discoverability and recognition, OpenOrgs combines automated processes with human curation. Numerous data sources are used to gather and merge information on organizations. Their metadata are automatically compared and combined, then these suggested identities are manually checked by data curators assigned at a national or multi-national level.

These two steps work as follows:

1. The deduplication algorithm (De Bonis, 2022) suggests a similarity between organizations that emerge in various sources by comparing their metadata (eg. the organization name, URL, country).

2. The automatic procedure is then verified by a manual curation process. By indicating whether two or more entities pertain to the same organization or not, data curators can clear up any ambiguity surrounding duplicates detected using the automated approach. Furthermore, they can themselves suggest new duplicates unidentified by the algorithm and make up for the information shortage by editing metadata description of the organization, compensating the lack of information from sources and enhancing the organization records' completeness and discoverability, for

example by adding a persistent identifier, an alternative name (OpenAIRE, 2023), or establishing parent- child relationships (eg. university and departments). As of now, there are more than 70 registered data curators from over 40 countries, with more than 100,000 curated organizations.

OpenOrgs offers a number of advantages for researchers, Research Performing Organizations (RPOs), Research Funding Organizations (RFOs), and all other stakeholders of Open Science services. It improves the findability of digital objects for academics and provides RPOs with a consistent showcase of the overall scientific production. It offers RFOs consistent data on the impact of resources. Finally, OpenOrgs offers functional and up-to-date services to all parties involved in Open Science.

In the OpenAIRE ecosystem, OpenOrgs plays an important role. For example, OpenAIRE Explore displays the curated metadata from OpenOrgs, giving researchers quick and easy access to details about the organizations involved in the research process (OpenAIRE EXPLORE, n.d.). These data are also used by the OpenAIRE Monitor service, which tracks and monitors research activities and Open Science trends of organizations (OpenAIRE MONITOR, n.d.). This integration improves these organizations' discoverability and recognition even more, fostering a more open and cooperative research environment. Therefore, rather than just a tool, OpenOrgs is a game-changer for the research community, and we believe it will contribute positively to build and maintain an integrated open scholarly communication system in the years to come.

KEYWORDS

data curation; deduplication; disambiguation; OpenAIRE; research organizations

REFERENCES

1. OpenOrgs: the OpenAIRE service for bridging registries of research organisations. (2023, May 24). OpenAIRE. *https://www.openaire.eu/openorgs-the-openaire-service-for-bridging- registries-of-research-organisations*

2. Pavone, G. (2021, October 27). OpenOrgs: Bridging registries of research organisations. OpenAIRE. *https://www.openaire.eu/blogs/openorgs-bridging-registries-of-research-organisations*

3. De Bonis, M.; Manghi, P.; Atzori, C. (2022). FDup: a framework for general-purpose and efficient entity deduplication of record collections. PeerJ Computer Science, 8:e1058 *https://doi.org/10.7717/peerj-cs.1058*

4. Artini, M.; La Bruzzo, S. F.; De Bonis, M.; Pavone, G. (2022). OpenOrgs: a tool for the disambiguation of organizations. ISNI Technical Reports, ISTI-2022-TR/034. Pisa: Istituto di Scienza e Tecnologie dell'Informazione. *https://bit.ly/3MkEO6b*

5. OpenAIRE Graph. (n.d.). Open. Transparent. Interconnected. *https://graph.openaire.eu/*

6. OpenAIRE EXPLORE. (n.d.). Discover open linked research. *https://explore.openaire.eu/*

7. OpenAIRE MONITOR. (n.d.). A new era of monitoring research. *https://monitor.openaire.eu/*

8. OpenAIRE. (n.d.). OpenOrgs Database. *https://orgs.openaire.eu/*