Data Search in practice: How to find scientific datasets and to link them to the literature

Ornella Irrera

Department of Information Engineering, University of Padova, Italy

Abstract

In the last years an increasing number of scientists, organizations and research communities started supporting Open Science principles and remarkably changed their research workflows; Open Science favors a more reliable and transparent science and supports the public access not only to the textual literature, but also to research data, code, and methodologies.

OpenAIRE is a European Project that has a key role in the promotion of Open Science principles providing a collection of infrastructures that aggregate metadata about research outputs and rely on them to populate a direct labeled graph, the OpenAIRE Research Graph, where literature is connected to the underlying research data. The lack of well defined and detailed metadata, prevents research data to be connected to other nodes in the graph, causing discoverability, reusability and reproducibility problems.

In this article we discuss a project with the purpose to effectively connect literature to research data in the OpenAIRE Research Graph relying on Data Search, Data Connection and Data Enrichment strategies.

Keywords

Open Science, Scholarly Communication, Data Search, OpenAIRE, OpenAIRE Research Graph

1. Introduction and Background

In the last decade, researching and sharing scientific results remarkably changed and more and more scientists recognized the importance and the necessity of supporting Open Science (OS) principles and goals. OS aims at making science more reliable and transparent; it promotes the public access to the scientific literature (*Open Access*), the sharing of scientific results, research materials, adopted methodologies (*Open Data*), and code and algorithms used throughout the whole research process (*Open Code*). This has positive implications for what concerns researchers and experiments. The public access to literature increases the number of citations, hence author's visibility: this poses the basis for an increase in collaborations between different research groups and an increase in multidisciplinarity. OS is central to favor the open access to data and it considerably changed scholarly communication by enhancing the role of data, giving credit to data creators, promoting the reusability of data and the reproducibility of the experiments [1, 2, 3]. One of the most important and active research infrastructures that promotes OS is OpenAIRE[1] (Open Access Infrastructure for Research in Europe) [4, 5].

FDIA2022: Future Directions in Information Access, July 20, 2022, Lisbon, Portugal
ornella.irrera@studenti.unipd.it (O. Irrera)
0000-0003-2284-5699 (O. Irrera)

[1] https://www.openaire.eu/

The purpose of the European OpenAIRE infrastructure is to facilitate, foster, support, and monitor OS and scholarly communication in Europe, promoting the adoption of global open standards and interoperability guidelines[2] to improve the scholarly communication ecosystem; in addition, it aims at providing a unique e-infrastructure for scientists to access open access publications [4, 6, 7]. To accomplish these objectives, OpenAIRE provides a set of technical infrastructures that aggregate metadata records of research products from a wide range of repositories and rely on these metadata to populate a direct labeled graph, the OpenAIRE Research Graph (ORG). The nodes of the ORG are the metadata records describing research outputs while the edges are links between metadata. Each edge has a label that identifies the semantics of the relationship (e.g. a publication *Cites* a dataset) [7].

What currently limits the full potential of the ORG as a mean to easily share and access research data, is the lack of detailed metadata describing research outputs other than the textual publication, such as datasets or software. According to [8], researchers still tend to share their data in many different ways: high variability in deposition approaches leads to the absence of well defined and standardized metadata; in the ORG, poor quality metadata prevent nodes to participate in a large number of links, hindering in this way the discoverability of the outputs and their authors.

In this article, we describe the key points of a system designed to find and connect literature to the underlying research data (the *datasets* in the ORG). The system is designed to work in a context where (i) datasets are not connected to the literature and (ii) metadata records describing datasets are not detailed enough to properly describe the output they represent. The proposed system is based on the integration of the following techniques:

- *Data Search*: definition of a set of methodologies to discover new datasets to connect to a publication (or vice versa).
- *Data Connection*: definition of a set of methodologies to create links between datasets and related publications, enhancing reproducibility and discoverability of experiments.
- *Data Enrichment*: definition of a set of methodologies that, given a dataset and/or some related papers, extract relevant information to populate useful descriptive metadata. We target the heterogeneity problem that affects the possibility of connecting the data to the literature.

The article is organized as follows. In section 2 we discuss the goals of the proposed system; in Section 3 we mention some approaches to link literature to the underlying data; in Section 4 we describe the ORG; in Section 5 we propose the key points of our approach.

2. Motivation and Objectives

The overall goals of this project are to develop a system able to find new relevant datasets, enrich the metadata records and connect datasets to one or more publications; the system is designed to run on the ORG (Section 4) provided by OpenAIRE. Currently, the largest part of the datasets in the graph lacks of connections with one or more publications: only 2 million datasets (out of 16 million research data currently searchable in OpenAIRE Explore[3]) are actually connected to

[2] https://guidelines.openaire.eu

[3] https://explore.openaire.eu/. These statistics refer to the Production version of the graph.

one or more publications. The lack of connections between publications and datasets nodes has important implications for what concerns *Discoverability*: datasets that are not linked to one or more publications, are represented as isolated nodes in the graph. This negatively affects:

- *Reusability* and *reproducibility*: problems in discoverability hinder the possibility to reuse the dataset to reproduce and perform new experiments.
- *Credit attribution*: the publication and the associated datasets may have disjointed sets of authors, and some authors may have contributed exclusively to the dataset instance. If a dataset is not easily discoverable, its contributors will not get the credits for that research output; in this regard, in [9] authors describe the problem of authorship, and propose an analysis focused on studying how many authors contributed to both the products connected by a *IsSupplementedBy* semantics, and how many of them have contributed exclusively to the publication or the research data.

The proposed system should give a valuable contribution in the OS domain, promoting datasets discoverability and reusability, improving the reproducibility of the scientific experiments and the credit attribution mechanisms.

3. Related Works

Connecting literature to the underlying research data is essential to effectively perform research communication. An important contribution in this context is provided by Scholix[4], an OpenAIRE service; it is a high level interoperability framework for exchanging information about the links between scholarly literature and data, as well as between datasets. It enables the communication between different links providers (such as OpenAIRE[5], Crossref[6] and DataCite[7]), providing a common information model [10].

In [11] it is discussed the cross-linking between journal publishers and data repositories for data publication. The first solution consists in relying on DOI as identifiers for the datasets: the dataset identifier is placed at the start of the paper and it is hyperlinked to the dataset landing page to be easily accessible. Another possible solution would be pulling metadata from the data repositories into Journal workflows and sharing the pre-publication metadata between data repositories and journals at the article submission stage: this ensures the consistency of information between data repositories and journals.

In order to make the datasets published on the Web easily accessible and discoverable, Google proposed a search engine for datasets, described in [12]. This search-engine is based on metadata records describing research data. As it was in OpenAIRE context, authors had to deal with complex problems such as the quality of metadata, the problem of dataset provenance (the same metadata can be hosted by different repositories), and the problem of the lack of a common metadata definition standard.

[4] http://www.scholix.org/
[5] https://www.openaire.eu/
[6] https://www.crossref.org/
[7] https://datacite.org/

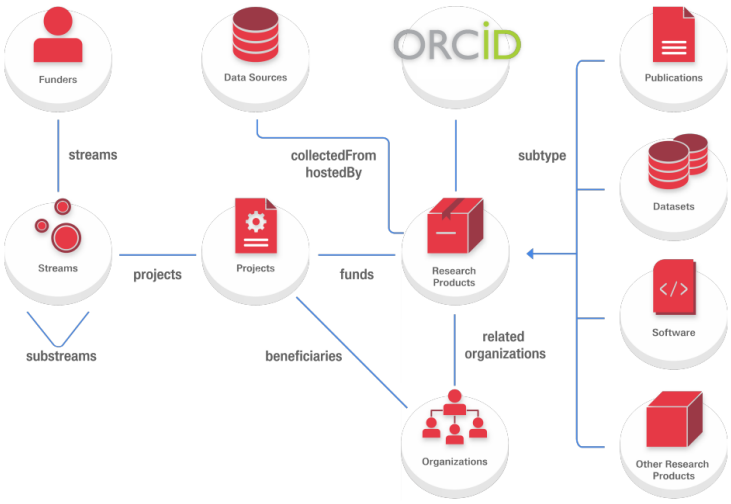

Figure 1: The ORG Graph Data Model [13].

4. The OpenAIRE Research Graph

To monitor research trends, interlink scientific results, keep track of provenance, promote a common scientific communication ecosystem, and access, share and reuse research outcomes, OpenAIRE provides the OpenAIRE Research Graph (ORG) [6, 14, 15, 16].

The ORG is a graph database where the metadata records representing products deriving from the research life-cycle are semantically interlinked [14, 17]. The structure and the semantics of the research graph are provided by the OpenAIRE Research Graph Data Model [6], depicted in Figure 1. According to the Graph Data Model, the graph can be described relying on three categories of entities:

- **Core entities**: they represent the nodes in the ORG. They can be *(i) Research results* such as publications, datasets or software; *(ii) Organization* intended as companies, research centers or institutions involved in projects; *(iii) Data Sources* such as publication repositories, datasets archives, funder databases which store the metadata; *(iv) Funders* that are the funders of projects that lead to the described results and that are responsible for one or more funding streams; *(v) Funding Stream* that is the investment from the funder; *(vi) Projects* are research projects funded by a Funding Stream.
- **Linking entities**: entities used to model relationships existing between the *Core entities*.
- **Types**: structured values for entity properties.

In order to generate the graph, OpenAIRE aggregates the metadata and links from 70 thousand trusted data sources. The metadata the scientists deposit about their research products (textual article, datasets, code, methodologies, and software) are put together with those derived from entity registries and they are finally transformed according to the OpenAIRE metadata model (*Aggregation*). The metadata corresponding to equivalent entities are merged (*Deduplication)* and inference algorithms are applied to metadata to infer new properties and links between

entities (*Enrichment*). The graph is finally cleaned in order to make the vocabulary compliant with the actual OpenAIRE controlled vocabulary and avoid possible errors that occurred in the previous phases [7, 14].

In the ORG it is possible to identify a set of *research communities*: these communities are portions of ORG focused on specific topics and research areas. It is possible to access the communities through the OpenAIRE Connect[8] infrastructure. The ORG is generated every two months on average and it is accessible via OpenAIRE Explore[9]. As of December 2021, the BETA version[10] of the graph counted more than 140 million publications, more than 50 million datasets, more than 258 thousand software and more than 3 billion of relationships. Relationships are bidirectional, hence each node interlinked with another one can be both *source* and *target* at the same time. The semantics of the relationships follow the DataCite Metadata schema[11].

5. Approach

In this section, we discuss the approach we propose to *(i) find isolated datasets to connect to the publications (and vice versa)* and *(ii) connect publications and datasets.*

The high heterogeneity and diversity of metadata records in the ORG allow us to subdivide our approach into four different research scenarios. The research scenarios are all focused on different portions of the ORG, extracted according to two criteria: the presence (or absence) of connections between publications and datasets nodes, and the level of detail which characterizes metadata. In the following paragraphs we call *rich* (or *detailed*) *metadata* the metadata records having the title, the description, the date of acceptance, the subjects (a list of keywords characterizing the resource topics) and the authors defined. Vice versa, research outputs missing one or more of these attributes are considered *poor*.

We propose an approach based on different scenarios in order to fragment the real-world heterogeneous data we are provided with into smaller parts: our intent is to isolate the data we are interested in and use them to develop data search, data connection and data enrichment techniques in different contexts; we start from the easiest condition and end with the real-world data.

Scenario 1 This scenario represents the *ideal* situation: the publications are semantically connected to the datasets and the metadata describing research outputs are detailed. The goal of this first scenario is the generation of a *ground-truth graph*: this is the gold standard we rely on to evaluate the techniques proposed in the other scenarios.

The satisfaction of the requirements above leads to consider a small portion of the original graph: the largest part of the graph, in fact, contains isolated datasets described by poor metadata. The subset of nodes and links considered to create the ground-truth belongs to the

[8] https://connect.openaire.eu/search/find/communities?type=%22community%22

[9] https://explore.openaire.eu/

[10] BETA version contains more nodes than the PRODUCTION version, which is the one available on OpenAIRE Explore. This is why in Section 2 we mentioned the presence of 16 million research data (less than 50 million found in BETA version).

[11] https://schema.datacite.org/meta/kernel-4.4/

European Marine Science (MES) community. As mentioned above, the constraints imposed in this scenario strongly affect the number of nodes used to generate the ground-truth. In the dump of December, 2021, in fact, the total number of datasets in MES is 118 thousand while those having rich metadata are 25 thousand. The initial count of MES publications instead, is 104 thousand, and those having rich metadata are 70 thousand. The requirement on the presence of connections between nodes further affects the total count of nodes and the final subset of nodes and links considered for the ground-truth graph counts 3967 publications, 5240 datasets, and less than 10 thousand semantic relationships.

Once extracted the subset of interest, we check the correctness of the metadata records and the semantic links provided by OpenAIRE; some metadata and relationships may be the result of inference algorithms and their correctness should be checked. To this aim, we consider not only the original set of metadata but also the textual article and the web pages of the datasets; these new sources of information have a dual role: to check if the metadata are correct and to enrich the current metadata set by adding new information such as sets of subjects and authors' ORCID and affiliation. The same sources of information are used to check the correctness of the semantic relationships: the information included in the textual documents and in the References section allows us to verify semantics such as *Cites* and *References*[12]. In addition, the web pages of the datasets usually contain information about the publications they are linked to; some data publishers specify also the semantics of the relationship (for example Zenodo[13]).

Finally, authors of research products are disambiguated. When scientists deposit information about authors, they can declare the same author in different ways in the publications deposition and in the dataset one (*O. Irrera* and *Ornella Irrera* are two ways to define the same person): in this case multiple definitions for the same author should be detected and removed.

The generation of the ground-truth allows us to apply in an ideal context *Data Connection* (we are able to detect the most probable semantics between two nodes) and *Data Enrichment* (we enrich the metadata thanks to the information extracted from different sources) strategies. The process of ground-truth generation poses the basis for the development of *Data Search* strategies: the analysis performed on the original graph in this first scenario eases the development of strategies to improve the discoverability of new datasets (or publications) to link.

Scenario 2 In this scenario, the publications and the datasets are connected but the metadata are poor. In this scenario we focus on *Data Enrichment* strategies.

Our starting point is a set of nodes connected but poorly described. We enrich the publications' metadata relying on the textual publication and we search for relevant information such as: abstract, title, authors, authors' ORCID, mail, and affiliations. In order to provide a set of relevant subjects about the article, we extract a set of entities from the article relying on *entity linking* methods, defined as methods designed to identify entities in text and link them to a knowledge repository [18]. This allows us to obtain a set of keywords that describe the articles' topics at best.

For what concerns datasets, we rely on different sources of information to enrich their metadata description. The first source is the dataset web page. In this web page, it is often possible

[12]The relationship starts from the publication and ends with the dataset: the publication *References* or *Cites* a dataset.
[13]https://zenodo.org/

to find the title, the description, the authors and other connected research outputs. Another source of information is the dataset itself; the dataset can provide relevant information such as data formats, results' units of measure, textual fields, columns names. Finally, the publication may contain relevant information about the dataset: textual articles in fact, sometimes discuss the data they rely on or produce.

The enriched sets of metadata allow us to develop *Data Connection* strategies and (i) check the correctness of the current semantics assigned to the edges and (ii) try to infer new edges. The developed methods are finally evaluated relying on the ground-truth generated in the previous scenario.

Scenario 3 In this scenario we examine a subgraph whose nodes have rich metadata but are not connected. This scenario is more complex than the previous ones; the absence of connections imposes the development of *Data Search* techniques to firstly discover, given a publication, a list of datasets which are probable to be semantically connected to it. Then, once obtained a list of candidates, we develop *Data Connection* techniques to predict the presence of connections and the related semantics.

Data Search techniques are based on specific metadata such as the title or the description: when a publication is deposited with its supplementary datasets, for example, they might share the entire description (which is the abstract of the article). In this case, the publication title can be contained in the dataset one (the title of the dataset can be: *Data for:* followed by the article's title). Basing on these two fields it is possible to find the nodes which are connected by a *IsSupplementedBy* semantics. Another possibility is to start from the publications' authors and search for all the research outputs an author is associated to; research outputs sharing the same authors are probable to be connected. The keywords associated to datasets and publications allow to cluster the nodes according to semantically similar subjects: in this way we isolate research outputs having common topics hence which are more probably connected. Finally, the information extracted from datasets' web pages provide relevant information about connected research outputs, especially when the semantics are specified.

Data Connection strategies allow us to predict the presence of a link between couples of nodes. This can be seen as a classification task where a machine learning model is trained to predict the presence of edges (and the related semantics) between couples of nodes represented by features vectors. To evaluate the prediction model, we rely on the ground-truth generated in the first scenario.

Scenario 4 The last scenario is the most complex one and it represents the *real* condition which affects the largest part of the graph. In this case, none of the requirements is satisfied: publications are not connected to datasets and metadata are poor. In this scenario, there is a continuous integration of *Data Search*, *Data Enrichment* and *Data Connection* strategies, which should not be intended as separated techniques to be applied one after to the other with a specific and predefined order. They should be considered, instead, as three integrated strategies which continuously cooperate to: *search* new datasets, *enrich* metadata and *connect* couples of nodes. A representation of this scenario is depicted in Figure 2.

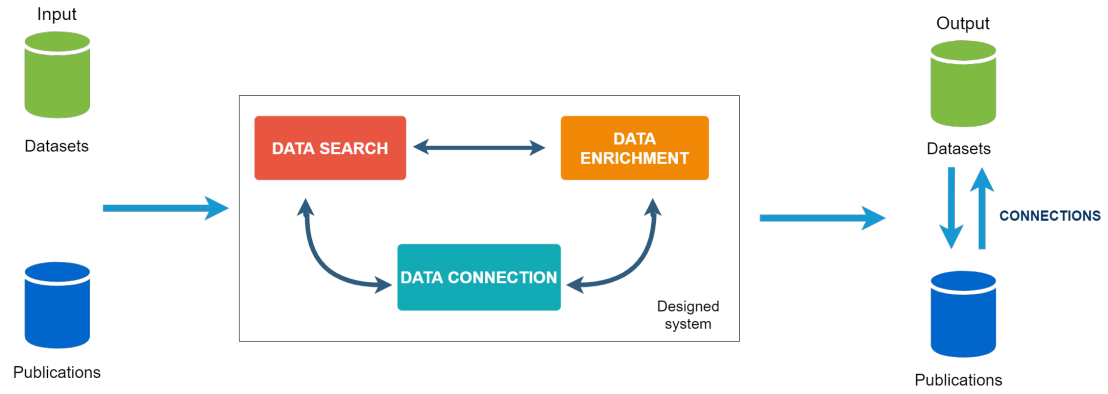

Figure 2: The cooperation of *Data Search*, *Data Connection*, *Data Enrichment* strategies is represented. The input of the system is a set of publications and datasets without links. The output is the set of publications semantically connected to one or more datasets and whose metadata are enriched.

6. Conclusions

In this paper we discussed a system to connect literature to the underlying research data in the ORG, a direct labeled graph where research outputs are semantically interlinked. Connecting research outputs such as literature, datasets and software is central to promote the discoverability of the outputs, methodologies, and data scientists relied on to perform the experiments. What actually hinders the possibility of creating new connections is the lack of metadata capable of describing research outputs in details: in the ORG these nodes are isolated and it is difficult to access them. The system we propose tackles the absence of connections and the incompleteness of metadata relying on the integration of: *Data Search* intended as a set of methodologies to easily discover new datasets (or publications) to connect; *Data Enrichment* intended as a set of methodologies to tackle the heterogeneity and incompleteness of metadata; *Data Connection* intended as a set of methodologies to find new links between literature and research data promoting the reproducibility and discoverability of the experiments. We proposed an approach based on four different research scenarios; each scenario has been defined according to two constraints: the presence of connections and the degree of completeness of the nodes' metadata; each scenario examines a portion of the graph extracted according to the constraints above. The first scenario represented the *ideal* condition. In this scenario we generate a ground-truth graph needed to evaluate the next three scenarios. The second and third scenarios are focused on the development of Data Search, Data Connection and Data Enrichment strategies. The fourth scenario represents the real condition, that is the condition that affects the largest part of the graph and the most difficult one. In this last scenario the combination of the techniques previously developed is applied. Our goal is the development of a system that can bring a valuable contribution to the OpenAIRE Project and, more in general in the OS context, providing a set of methodologies to create new connections between research outputs despite the absence of full representative metadata records.

Acknowledgments

This work was supported by the ExaMode Project, as a part of the European Union Horizon 2020 Program under grant 825292.
Thanks to Andrea Mannocci[14], Paolo Manghi[15] and Gianmaria Silvello[16].

References

[1] C. Allen, D. M. Mehler, Open science challenges, benefits and tips in early career and beyond, PLoS biology 17 (2019) e3000246.
[2] E. C. McKiernan, P. E. Bourne, C. T. Brown, S. Buck, A. Kenall, J. Lin, D. McDougall, B. A. Nosek, K. Ram, C. K. Soderberg, et al., Point of view: How open science helps researchers succeed, elife 5 (2016) e16800.
[3] National Academies of Sciences, Engineering, and Medicine and others, Open science by design: Realizing a vision for 21st century research (2018).
[4] N. Rettberg, B. Schmidt, Openaire—building a collaborative open access infrastructure for european researchers., Liber Quarterly: The Journal of European Research Libraries 22 (2012).
[5] P. Manghi, L. Bolikowski, N. Manola, J. Schirrwagen, T. Smith, Openaireplus: the european scholarly communication data infrastructure, D-Lib Magazine 18 (2012).
[6] P. Manghi, A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen, P. Principe, M. Artini, A. Becker, M. De Bonis, et al., The openaire research graph data model, Zenodo (2019).
[7] M. Baglioni, A. Bardi, A. Kokogiannaki, P. Manghi, K. Iatropoulou, P. Principe, A. Vieira, L. H. Nielsen, H. Dimitropoulos, I. Foufoulas, et al., The openaire research community dashboard: on blending scientific workflows and scientific publishing, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2019, pp. 56–69.
[8] I. J. Aalbersberg, J. Dunham, H. Koers, Connecting scientific articles with research data: New directions in online scholarly publishing, Data Science Journal 12 (2013) WDS235–WDS242.
[9] A. Mannocci, O. Irrera, P. Manghi, Will open science change authorship for good? towards a quantitative analysis, in: IRCDL, 2022.
[10] A. Burton, H. Koers, P. Manghi, M. Stocker, M. Fenner, A. Aryani, S. La Bruzzo, M. Diepenbroek, U. Schindler, The scholix framework for interoperability in data-literature information exchange, D-Lib Magazine 23 (2017).
[11] S. Callaghan, J. Tedds, R. Lawrence, F. Murphy, T. Roberts, W. Wilcox, Cross-linking between journal publications and data repositories: A selection of examples (2014).
[12] D. Brickley, M. Burgess, N. Noy, Google dataset search: Building a search engine for datasets in an open web ecosystem, in: The World Wide Web Conference, 2019, pp. 1365–1375.
[13] K. Vichos, M. De Bonis, I. Kanellos, S. Chatzopoulos, C. Atzori, N. Manola, P. Manghi,

[14]Andrea Mannocci, 0000-0002-5193-7851
[15]Paolo Manghi, 0000-0001-7291-3210
[16]Gianmaria Silvello, 0000-0003-4970-4554

T. Vergoulis, A preliminary assessment of the article deduplication algorithm used for the openaire research graph, in: IRCDL, 2022.
[14] P. Manghi, https://www.openaire.eu/blogs/the-openaire-research-graph, 2019. Accessed: 2022-06-06.
[15] https://www.openaire.eu/aggregation-and-content-provision-workflows, 2022. Accessed: 2022-06-06.
[16] A. Pavlidou, https://www.openaire.eu/blogs/openaire-research-graph-an-intelligent-gateway-to-scholarly-communication, 2021. Accessed: 2022-06-06.
[17] https://graph.openaire.eu/about, 2019. Accessed: 2022-06-06.
[18] K. Balog, Entity linking, in: Entity-Oriented Search, Springer, 2018, pp. 147–188.