D-Lib Magazine

November/December 2015
Volume 21, Number 11/12

## The OpenAIRE Literature Broker Service for Institutional Repositories

Michele Artini, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi and Andrea Mannocci
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" — CNR, Pisa, Italy
{michele.artini, claudio.atzori, alessia.bardi, sandro.labruzzo, paolo.manghi, andrea.mannocci}@isti.cnr.it

DOI: 10.1045/november2015-artini

### Abstract

OpenAIRE is the European infrastructure for Open Access scholarly communication. It populates and provides access to a graph of objects relative to publications, datasets, people, organizations, projects, and funders aggregated from a variety of data sources, such as institutional repositories, data archives, journals, and CRIS systems. Thanks to infrastructure services, objects in the graph are harmonized to achieve semantic homogeneity, de-duplicated to avoid ambiguities, and enriched with missing properties and/or relationships. OpenAIRE data sources interested in enhancing or incrementing their content may benefit in a number of ways from this graph. This paper presents the high-level architecture behind the realization of an institutional repository Literature Broker Service for OpenAIRE. The Service implements a subscription and notification paradigm supporting institutional repositories willing to: (i) learn about publication objects in OpenAIRE that do not appear in their collection but may be pertinent to it, and (ii) learn about extra properties or relationships relative to publication objects in their collection.

## 1 Introduction

The OpenAIRE infrastructure [5] is both a networking and technological infrastructure whose mission is to advocate and monitor the adoption of the European Commission Open Access mandates, and to evaluate the impact of EC funding and National funders.

Its networking infrastructure consists of the National Open Access Desks (NOADs), providing OpenAIRE contact points for each of the EC countries. The NOADs monitor and advocate the adoption of Open Access EC policies at the level of the countries, support researchers at the implementation of the EC Data Pilot, and function as a bidirectional communication channel between the Commission, OpenAIRE and the countries.

Its technological infrastructure provides services [6] to monitor funders, project research impact, and track Open Access trends in terms of related publications and datasets. To this aim, the services offer functionalities to populate a European (and beyond) graph-like information space that aggregates information about publications, datasets, organizations, persons, projects and several funders (e.g. European Commission, Wellcome Trust, Fundação para a Ciência e a Tecnologia, Australian Research Council) collected from hundreds of online data sources (e.g. publication repositories, dataset repositories, and CRIS systems, journals, publishers). To facilitate the harvesting process as well as interoperability between publication repositories, dataset repositories, and CRIS systems, OpenAIRE has released specific "metadata export guidelines" for the managers of such data sources [1][8]. The guidelines describe the expected structure (i.e. fields) and semantics (i.e. vocabularies and formats) of the metadata records, as they should be exposed by the data sources. Their aim is to reach a community consensus on how to homogenize and therefore facilitate exchange of information across scholarly communication data sources in Europe (and beyond, thanks to the synergies with COAR, US SHARE [3], and UK JISC). The typologies and the number of data sources currently included in OpenAIRE are summarized in Table 1.

| Data Source Typology | Number of Data Sources | Type of Objects |
|---|---|---|
| Journal Platform | 5,582 | Publications and persons |
| Publication Repository | 512 (Total) | Publications and persons |
| Institutional | 426 | |
| Thematic | 36 | |
| Other/Unknown | 50 | |
| Data Repository | 38 | Datasets, publications, and persons |
| Aggregator of Publication Repositories | 8 | Publications and persons |
| Aggregator of Data Repositories | 1 | Datasets and persons |

| | | |
|---|---|---|
| Aggregator/Publisher of Journals | 6 | Publications, persons, and data sources (i.e. journals) |
| Entity Registry (data sources offering authoritative lists of entities) | 13 | Data sources (i.e. publication repositories, data repositories), projects, funders, persons |
| CRIS systems | 0 (the first CRIS systems will be aggregated by the end of 2015) | Publications, datasets, projects, persons |

Table 1: Data source typologies in the OpenAIRE federation (update to date 2015-10-16)

The OpenAIRE infrastructure collects metadata records from data sources and derives from them objects and relationships that form the information space graph. For example, a bibliographic metadata record describing a scientific article will yield one publication object and a set of person objects (one per author) related to it. After aggregation, dedicated services clean and enrich the graph as depicted in Figure 1:

- Harmonization (aggregation sub-system): objects of given entities are transformed from their native data models (e.g. physically represented as XML records, HTML responses, CSV files) onto the OpenAIRE data model [7] in order to build an homogenous information space.
- Merge (de-duplication subsystem): objects of the same entity type are de-duplicated in order to remove ambiguities that may compromise statistics (e.g. the same publication may be collected from different repositories as supposedly different objects).
- Enrichment (information inference sub-system): publication full-texts are collected and processed by text mining services [9] capable of inferring new property values or new relationships between objects.

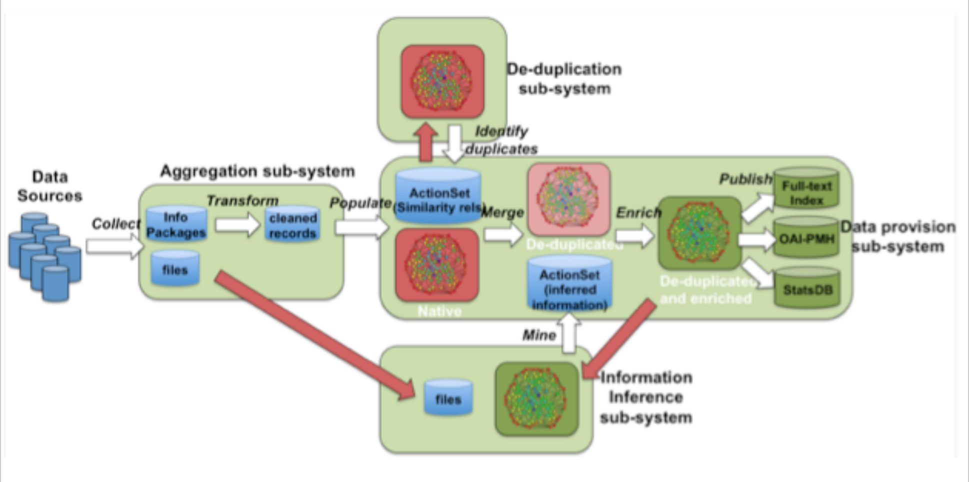

*Figure 1: OpenAIRE services high-level architecture*

The enriched information space graph is then made available for programmatic access via several APIs (Search HTTP APIs, OAI-PMH, and soon Linked Open Data) [2] and for search, browse and statistics consultation via the OpenAIRE portal.

Needless to say, data sources that are providing content to OpenAIRE and are interested in augmenting their local collections may benefit in a number of ways from the OpenAIRE information space. This is particularly true for institutional repositories, whose mission is to grow a complete collection of the scientific publications produced by the authors affiliated with the institution they serve. The repository managers' goal is twofold: to bring into the collection all articles produced by affiliated authors, and to make sure that the metadata is as complete and up-to-date as possible.

This paper presents the functional requirements driving the realization of a Literature Broker Service for the OpenAIRE infrastructure. The Service implements a subscription and notification mechanism supporting repository managers who are enhancing the content of their repositories by taking advantage of the OpenAIRE information space. Using the Service, repository managers can subscribe to special "addition" or "enrichment" events in order to be notified about: (i) publication objects in OpenAIRE that do not appear in their collection but may be pertinent to it, or (ii) properties or relationships relative to publication objects in their collection that do not appear in their local metadata.

Section 2 describes two initiatives for the brokerage of publication metadata: the US SHARE Notify and the JISC/EDINA Publications Router. Section 3 presents the OpenAIRE graph data model and its approach to modelling the provenance of the original metadata records and of the inferred properties and relationships. The opportunities of data exchange between the OpenAIRE infrastructure and institutional repositories are discussed in Section 4, where the OpenAIRE Broker Service and its subscription and notification mechanisms are also presented. Section 5 offers conclusions and discusses future work.

## 2 Repository Literature Brokers in the literature

The literature deluge makes the reporting and tracking of research results harder for all stakeholders in scholarly communication. Researchers often feel they lose precious time when they are asked to provide detailed metadata information about their articles multiple times at different locations, e.g. the institutional repository and funders. As a consequence, publication metadata can be poor, subject to mistakes, and found at different locations. Publishers own publication metadata information, but a direct interaction with repositories is rare, for both technical (e.g. lack of shared author identifiers) and cost reasons. As a consequence, a number of initiatives started working on approaches favoring single-deposition of publication metadata with subsequent automated delivery to other repositories. Some approaches focused on techniques for automatic deposition into a repository (SWORD project [4]), while others focused on the complementary aspects of how to broker publication information from publishers to relevant/interested repositories. SHARE and JISC/EDINA are two such initiatives, based respectively in the US and UK

SHARE (SHared Access Research Ecosystem) [3] is a higher education and research community established in 2013 that supports preservation, access and re-use of research results across United States. The first project set up by SHARE is SHARE Notify. A public beta version of the service has been available since April 2015 and counts approximately 600,000 metadata records about articles and datasets from more than 30 providers. SHARE Notify allows interested stakeholders (e.g. researchers, repositories, funders) to subscribe to notifications about research release events such as the publication of an article in a peer-reviewed journal, the deposition of a pre-print version in an institutional repository or the deposition of a dataset. Notifications are distributed as Atom feeds, consumable via common RSS readers, containing metadata summaries about the research results matching the subscription query. The subscription query may include any metadata field of the SHARE schema, also in combination with boolean operators according to the Lucene syntax. While it is possible to subscribe to be notified of events related to one or more data sources (journals, repositories, etc.), it is not yet possible to subscribe to receive events related to authors' institutions, as the majority of metadata records collected by SHARE does not contain explicit authors' affiliations. A JSON API is also available to build dedicated applications by consuming the content collected by SHARE.

JISC is a UK initiative that promotes ICT in education and research. Built in collaboration with EDINA, the prototype of the JISC Publications Router offers a notification system (PostCards) and an automatic mechanism based on the SWORD protocol to transfer metadata and files from one location to another. Upon subscription, users can select one or more repositories of interest. The Postcards system will send them emails with a list of metadata records suitable for the selected repositories. The "suitability" of a record with respect to a given repository is automatically calculated by extracting the authors' affiliations from the metadata records. Though the subscription criteria are static, the Postcard system of the JISC Publications Router is very flexible in terms of the format of the notifications: citations in ASCII, bibtex, endnote, and Dublin Core metadata records are only a subset of the formats that a subscriber can choose to receive. Currently, the service is undergoing a full revision to improve its quality and make it a production system in 2016 [10].

The JISC/EDINA Publications Router is more mature than SHARE Notify and its capabilities of detecting authors' affiliations and of sending notifications in different formats are valuable. Metadata records collected by the router can be bulk downloaded via the standard OAI-PMH protocol. On the other hand, the "young" SHARE Notify gives more control to users on their subscription topics and the availability of a JSON API allows IT-skilled users to build applications on top of the SHARE content.

## 3 OpenAIRE information space

The OpenAIRE information space data model [7] (see Figure 2) builds on the OpenAIRE guidelines and is inspired by the DataCite and CERIF initiatives. Its main entities are: *Results* (datasets and publications), *Persons*, *Organizations*, *Funders*, *Funding Streams*, *Projects*, and *Data Sources*.

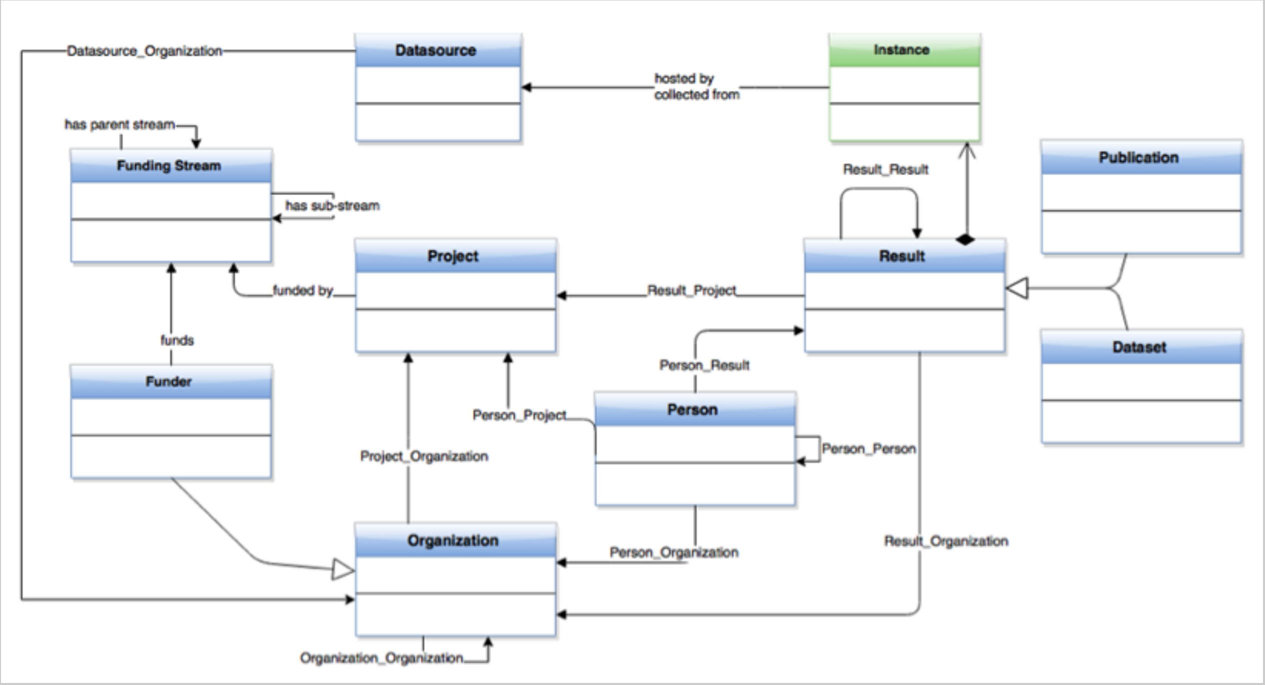

*Figure 2: The OpenAIRE data model*

*Results* are intended as the outcome of research activities and may be related to *Projects*. OpenAIRE supports two kinds of research outcome: *Datasets* (e.g. experimental data) and *Publications* (*Patents* and *Software* entity types will be introduced soon). As a result of merging equivalent objects collected from separate data sources, a Result object may have several physical manifestations, called *instances*; instances indicate URL(s) of the payload file, access rights (i.e. open, embargo, restricted, closed), and a relationship to the data source that hosts the file (i.e. provenance).

*Persons* are individuals that have one (or more) role(s) in the research domain, such as authors of a Result or coordinator of a Project.

*Organizations* include companies, research centers or institutions involved as project partners or that are responsible for operating data sources.

*Funders* (e.g. European Commission, Wellcome Trust, FCT Portugal, Australian Research Council) are *Organizations* responsible for a list of Funding Streams (e.g. FP7 and H2020 for the EC), which are strands of investments.

*Funding Streams* identify the strands of funding managed by a *Funder* and can be nested to form a tree of sub-funding streams (e.g. FP7 — SP1 — HEALTH).

*Projects* are research projects funded by a *Funding Stream* of a *Funder*. Investigations and studies conducted in the context of a *Project* may lead to one or more *Results*.

Finally, OpenAIRE objects are created out of metadata records (e.g. XMLs, CSV, txt, xls, JSON, HTML) collected from various *Data Sources* (see Table 1). Data Sources are associated with all objects collected from them.

In order to give visibility to the original data sources, OpenAIRE keeps provenance information about each piece of aggregated information. Specifically, since de-duplication merges objects collected from different sources and inference enriches such objects, provenance information is kept at the granularity of the object itself, its properties, and its relationships. Object level provenance tells the origin of the object that is the data sources from which its different manifestations were collected. Property and relationship level provenance tells the origin of a specific property or relationship when inference algorithms derive these, e.g. algorithm name and version. Examples are:

- Document classification properties: e.g. subjects from a set of standard classification schemes, such as the Dewey Decimal Classification and Medical Subject Headings;
- Research initiative properties: e.g. information about the research initiatives, such as the European Grid Infrastructure, related to the research results presented in the publication;
- Citation properties: e.g. the list of references cited by the publication, extracted from the bibliography or reference section of the full-text;
- Relationships to projects, datasets, and similar publications.

## 4 The OpenAIRE Literature Broker Service

The OpenAIRE enriched information graph offers a great opportunity for managers of institutional repositories to improve their collections. However, the current access APIs provided by OpenAIRE to third-party services (i.e. HTTP APIs, OAI-PMH) are not intended to support these needs. To this aim, the infrastructure is in the process of realizing a Literature Broker Service ("the Service"), by learning from other experiences, and targeting the specificity of the OpenAIRE setting. The Service will allow repository managers to subscribe to (potential) "enrichment" and (potential) "addition" events occurring to the OpenAIRE information space graph with respect to the scope of their repository. "Enrichment" events identify objects fed by the repository to OpenAIRE that have been enriched by OpenAIRE inference algorithms or de-duplication merges (i.e. merged with objects describing the same publication but with richer or different metadata, for example the open access version of an article). "Addition" events identify objects that enter into the OpenAIRE information space graph, are not present in the repository, but may be part of its collection. Repository managers will then receive notifications about the events they are subscribed to, according to various notification strategies.

Figure 3 shows how the Service will integrate with the existing OpenAIRE infrastructure. Objects collected from data sources are aggregated, de-duplicated and enriched by inference algorithms to form the OpenAIRE Information space graph. Whenever a new information space is generated, the Service explores the graph to detect if any of the active subscriptions are matched and if so, the active subscriptions are matched and if so, notifications are generated and delivered.

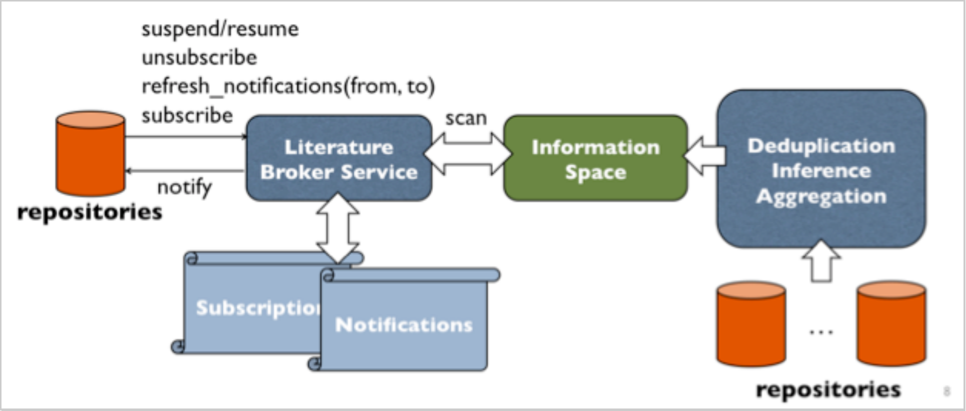

*Figure 3: The Scholarly Communication Broker Service in the OpenAIRE infrastructure*

### 4.1 Subscriptions

Repository managers will be able to subscribe to two main classes of subscriptions: "enrichment" and "addition".

The first class refers to notifications about publications that (i) were collected from the repository by OpenAIRE and (ii) have been enriched with properties or relationships to other objects by OpenAIRE inference algorithms (e.g. relationships to projects and datasets, citation lists, document classification properties) or by the side effect of being merged with richer publication objects (e.g. DOI of a publication, Open Access version of the publication). The identification of these events is straightforward as it is based on provenance of collection (i.e. selects publication objects collected from the given repository) and of enrichment (i.e. further selects objects of the given repository involved into a merge or enriched by inference algorithms). Repository managers will be able to fine-tune their subscriptions based on the typology of enrichment.

The second class refers to notifications relative to publications that are "relevant to" the repository at hand, but are not present in the repository. The identification of these events is less trivial as it requires devising a criterion of "publication *relevant to* a repository". Three strategies have been proposed, according to which a publication is relevant to a repository if one of the following chains of relationships exist in the OpenAIRE information graph:

- *publication-author-organization-repository*: the publication has an author whose organization (affiliation) has a given institutional repository of reference;
- *publication-author-repository*: the publication has an author with a given institutional repository of reference;
- *publication-project-organization-repository*: the publication has been funded by a project whose participants (beneficiaries of the grant) have a given institutional repository of reference.

Given a publication, if such relationships exist as are found in the graph (collected or inferred) the Service may notify the interested repositories of the publication. The challenge is that such relationships are generally not provided by data sources but must be inferred by OpenAIRE services. As a consequence subscription and notification can secure levels of "correctness" that depend on the level of trust of inference algorithms and can be fine-tuned by repository managers at subscription time.

***Relationships: publication-author-organization-repository***

The most intuitive criterion of publication *relevant to* a repository is that based on the relationships *authorship*, i.e. the publication has a given author, *author affiliation*, i.e. the author of the publication is affiliated with an organization, and *organizationRepositoryOfReference*, i.e. the institutional repository of reference of all authors of an organization. While OpenAIRE can collect from data sources relationships between publication-author (e.g. publication metadata) and data source-organizations (e.g. OpenDOAR returns the list of European publication repositories), affiliational relationships between publication and authors are generally not available in collected publication metadata. In fact, publication records provided by data sources (according to the OpenAIRE guidelines, but also according to accepted use of Dublin Core) do not provide author affiliation information or when they do, they follow patterns that vary from case to case and are hard to match automatically.

The inference service of OpenAIRE features a module for affiliation inference, which mines the publication full-texts to identify and extract pairs <author-organization>. If the algorithm is able to determine which author is associated with which organization, then a relationship *affiliation* between the author and the organization is added to the graph, otherwise the *authoringOrganization* relationship is created between the publication and the organization. For the purpose of the Service there is no difference between the two (see Figure 4) as what matters is the identification of a relationship between the publication and repositories.

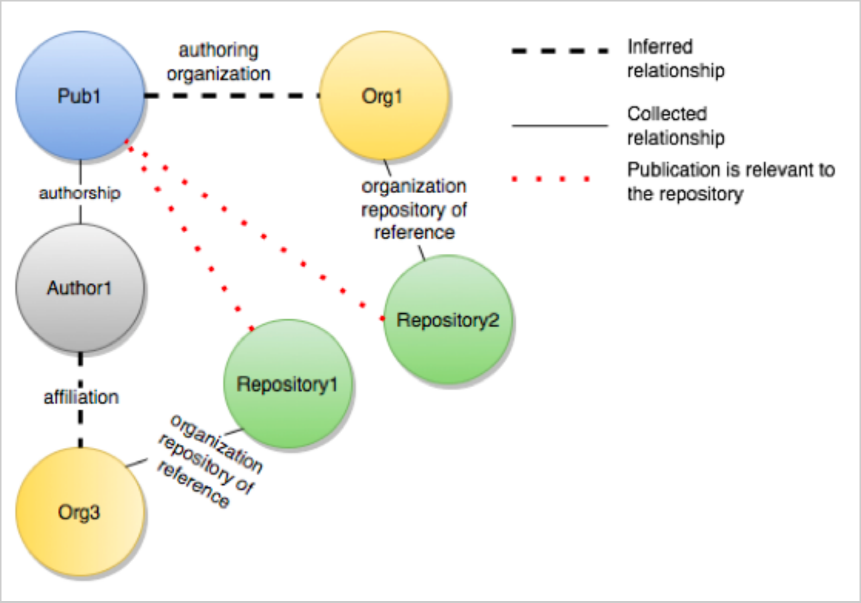

*Figure 4: Detection of "relevant to" criterion via full-text mining*

***Relationships: publication-author-repository***

The second criterion, to detect which publications may be "relevant to" a repository, is based on the relationships *authorship*, i.e. the publication has an author, and *authorRepositoryOfReference*, the author deposits her publication in the given repository. As previously mentioned, *authorship* is generally provided by the collected metadata, while *authorRepositoryOfReference* needs to be inferred by OpenAIRE services. To this aim the services exploit the results of the de-duplication algorithms over authors and publications. Harvested metadata records contain authors' names in *dc:creator* fields, as simple strings. In the OpenAIRE information space, such "raw" author objects are initially created with a stateless identifier that makes them unique in the graph. Author identifiers are obtained from the OpenAIRE identifier of the repository, the OpenAIRE identifier of the publication that contains them (obtained from publication identifiers such as DOIs or OAI-PMH identifiers), and the author name string (see Figure 5 for an example).

As such, before de-duplication, each occurrence of an author name in a publication from a given data source is considered to be a unique author, which carries a pointer to the data source (e.g. the repository) and to the publication that brought it into the system. The result of de-duplication over author objects is a set of "anchor" authors obtained as the merge of several "raw" authors. Starting from "anchor" authors, and exploiting the pointers to institutional repositories of the raw authors they merge, OpenAIRE inference services calculate the notion of "author submission frequency" by counting the number of publications of the author across different repository data sources. In the majority of cases, the repository with the highest submission frequency turns out to be the repository of reference for the author, namely the one to which she is supposed to report her publications (in future work, this process will be further refined to identify the "migration" of an author to another institution, therefore to a different repository of reference; this condition may conflict with the "highest number of submissions" criteria, but may be identified using submission dates).

Accordingly, with a given degree of approximation, when OpenAIRE collects a new publication from a given repository it is possible to state if some of its authors have a different repository of reference. An exemplification is shown in Figure 5. The same author string ("A. Turing") collected from four different data sources (three institutional repositories and one data source of a different typology) results in four different person objects. When the de-duplication is run, the four persons are merged into one new "anchor" object ("anchor::A. Turing", in the example). Table 2 shows the occurrences of submission of "anchor::A. Turing" considering the provenance of the four "raw" authors it merges. The table shows that the author deposited mostly in repository "Repo1", which can then be considered the repository of reference for the author. Consequently, "Repo1" may be interested in being notified about the publications the author deposited in "Repo2", "Repo3" and "DS".

In order to avoid useless notifications, publication de-duplication permits understanding whether or not the repository of reference already has the publications or should be notified.

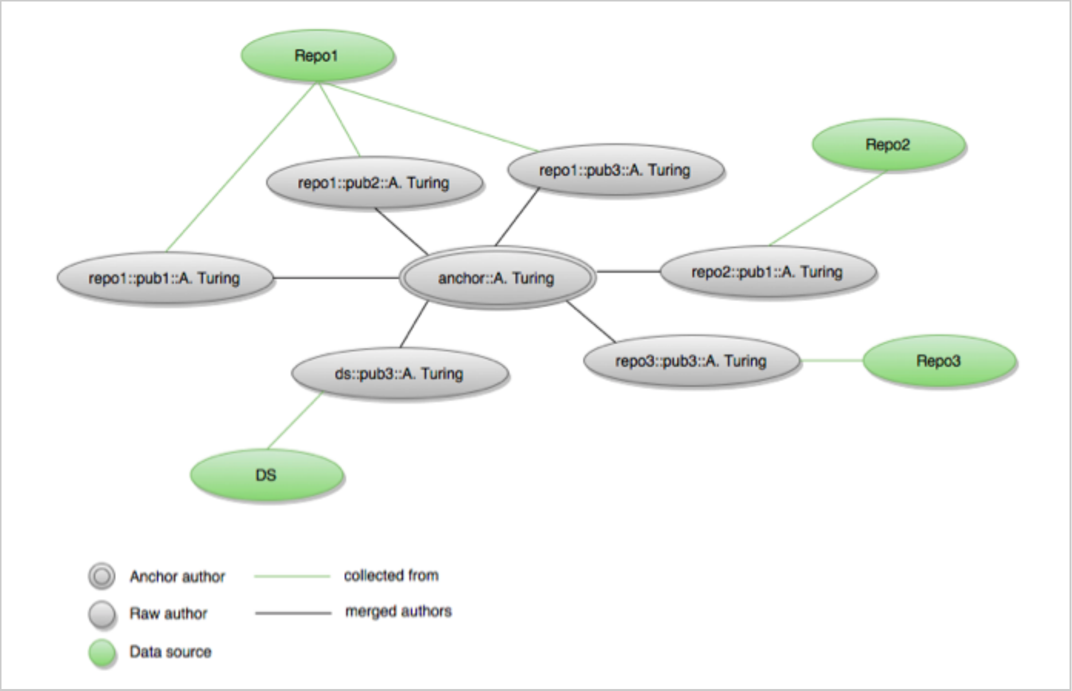

*Figure 5: Affiliation detection: using de-duplication to compute the closeness of an author to a repository*

| | **Repo1** | **Repo2** | **Repo3** | **DS** |
|---|---|---|---|---|
| Anchor::A. Turing | 3 | 1 | 1 | 1 |

Table 2: Submission frequency for the graph in Figure 5

The accuracy of the de-duplication algorithms is very important for the correct implementation of this strategy. In addition, repository managers should be able to fine-tune the parameters for the selection of "author submission frequency" (e.g. minimum number of submissions per data source or in total) in order to limit the number of false positive notifications.

A preliminary analysis of the OpenAIRE information space graph for the detection of "frequent submitters" has been carried out considering authors with at least 10 publications and with at least 4 publications in the repository of reference. The analysis is summarized in Table 3 and in the graph in Figure 6. From a total of 157,549 anchor authors from 426 institutional repositories, about 19% submit their articles into a single repository (i.e. 100% of the publications of each author has been collected from the same data source). Interestingly, about 60% submitted publications in different repositories, but their repository of reference hosts from 50% to 99% of their publications. Most likely, repository managers will be interested in this subset of authors, as they are those that mostly deposit papers in one repository, but some of their papers can also be found in other locations. Finally, about 20% submitted only up to 50% of their publications in their repository of reference.

| Author publications appearing in repository of reference (%) | Number of authors in this category |
|---|---|
| 10-19% | 70 |
| 20-29% | 4,469 |
| 30-39% | 12,316 |
| 40-49% | 15,802 |
| 50-59% | 20,659 |
| 60-69% | 16,832 |
| 70-79% | 15,805 |
| 80-89% | 18,823 |
| 90-99% | 22,572 |
| 100% | 30,201 |

Table 3: Authors, repositories of reference and percentage of deposition

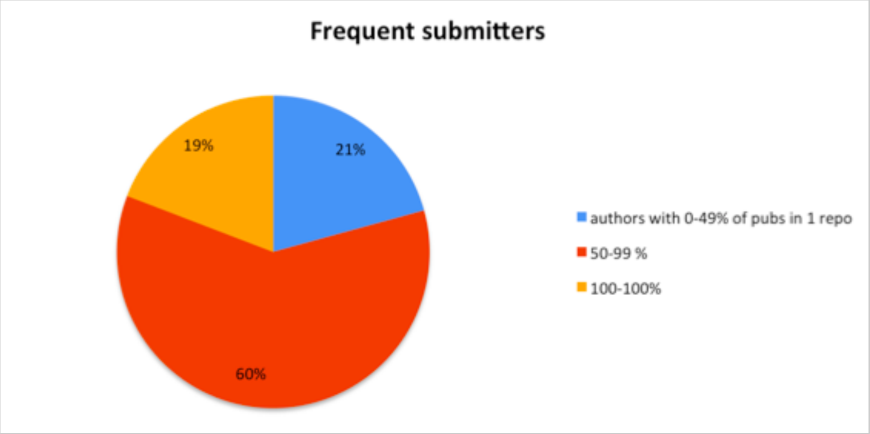

*Figure 6: Preliminary analysis of the OpenAIRE graph for "frequent submitters"*

***Relationships: publication-project-organization-repository***

The third criterion available for subscriptions on "relevant to" exploits the relationships *beneficiaryOf*, i.e. organizations involved in research projects, and *organizationRepositoryOfReference*, i.e. the institutional repository of reference for all authors of an organization. OpenAIRE collects these relationships from publication metadata (e.g. repositories, journals), project metadata (funders), and repository metadata (OpenDOAR). Figure 7 provides an example of the concept: CNR-ISTI is an Italian research institute whose institutional repository is PUblication MAnagement (PUMA). Researchers from CNR-ISTI should deposit their publications in PUMA. CNR-ISTI is involved in the EC funded project OpenAIRE2020, therefore some of the publications linked to the OpenAIRE2020 project may be "relevant to" PUMA because they may be co-authored by researchers working at CNR-ISTI.

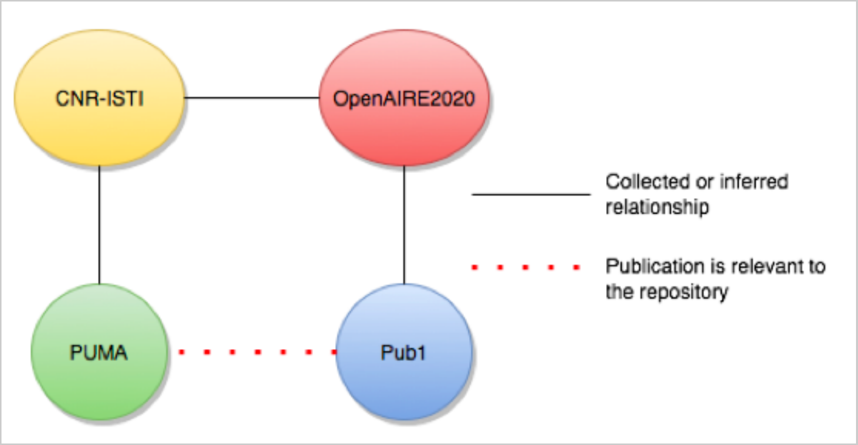
CNR-ISTI
OpenAIRE2020
Collected or inferred relationship
Publication is relevant to the repository
PUMA
Pub1

*Figure 7: Detection of publications' affiliation: exploiting links to projects*

The approach has a high chance of yielding false positive notifications because some projects involve a considerable number of organizations (e.g. the OpenAIRE2020 EC-H2020 project involves 49 organizations). Repository managers will be able to fine-tune their subscription in order to include, for example, only projects from a given list, or projects with a limited number of participants.

### 4.2 Notifications

Different notification strategies are under evaluation in order to meet the diverse requirements of subscribers:

- **Mail postcards**: Following the example of the JISC Publication Router, subscribers may opt to be notified by email at given intervals (e.g. daily, weekly, monthly) and with given granularity (individual records, digests, URL to user interface), together with instructions for the retrieval of the complete metadata records and full-texts.
- **Programmatic access**: APIs will be provided to retrieve notifications by status (e.g. read/unread), subscription typology, and filters (e.g. criteria on the metadata fields). A prototype solution based on OAI-PMH has already been realized on top of the OpenAIRE OAI-PMH Publisher and it is currently undergoing testing. For subscribing repositories, the OAI-PMH publisher service gives access to the OAI set of records collected from the repository that have been enriched by OpenAIRE.
- **Web interface**: A web application will offer a dashboard where repository managers can find the tools to:
  - Manage their notifications, i.e. create, suspend, resume and delete
  - View, download or re-send old notifications
  - Select the format of the email notification digests to receive among a set of supported formats (e.g. Dublin Core XML, Bibtex and citations in ASCII)

Finally, existing APIs for the automatic ingestion of records into repositories will be evaluated (SWORD, [10]) and realization of software modules for integration and ingestion into known repository platforms will be considered (e.g. DSpace, ePrints).

## 5 Conclusions

OpenAIRE populates, cleans, and enriches a graph of objects relative to publications, datasets, people, organizations, projects, and funders aggregated from a variety of data sources. The OpenAIRE graph is a great opportunity for repository managers to improve their repository collections, as it may feature information that is not otherwise available to them. The OpenAIRE Literature Broker Service will offer subscription and notification functionalities explicitly targeting their needs. By exploiting the provenance information tracked by the OpenAIRE infrastructure, it will be possible to subscribe to "enrichment" events and be notified whenever OpenAIRE enriches a publication metadata record with new properties (subjects, citation list, research initiatives) or new relationships to projects or datasets. By enriching with relationships and analyzing the information space graph, the service will also be able to notify repository managers about "addition" events whenever a publication metadata record relevant for that repository is aggregated from another data source.

A first prototype for the export of enriched metadata records per data source via OAI-PMH has already been implemented, and in the near future, the procedure for subscription and email notification, together with a first implementation of the Dashboard Web UI, will be made available to selected repository managers for testing purposes.

## Acknowledgements

This work is partially funded by the European Commission H2020 project OpenAIRE2020 (Grant agreement: 643410, Call: H2020-EINFRA-2014-1).

## References

[1] The OpenAIRE guidelines.

[2] The OpenAIRE API.

[3] Walters, T., & Ruttenberg, J. (2014). SHared Access Research Ecosystem. *Educause Review*, 49(2), 56-57.

[4] Lewis, S., de Castro, P., & Jones, R. (2012). SWORD: Facilitating deposit scenarios. *D-Lib Magazine*, 18(1), 4. http://doi.org/10.1045/january2012-lewis

[5] Manghi, P., Bolikowski, L., Manold, N., Schirrwagen, J., & Smith, T. (2012). Openaireplus: the european scholarly communication data infrastructure. *D-Lib Magazine*, 18(9), 1. http://doi.org/10.1045/september2012-manghi

[6] Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela L., Castelli D., & Pagano P. (2014). The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program: electronic library and information systems*, 48(4), 322-354. http://doi.org/10.1108/PROG-08-2013-0045

[7] Manghi, P., Houssos, N., Mikulicic, M., & Jörg, B. (2012). The data model of the openaire scientific communication e-infrastructure. In *Metadata and Semantics Research* (pp. 168-180). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-35233-1_18

[8] Houssos, N., Jörg, B., Dvořák, J., Príncipe, P., Rodrigues, E., Manghi, P., & Elbæk, M. K. (2014). OpenAIRE guidelines for CRIS managers: supporting interoperability of open research information through established standards. *Procedia Computer Science*, 33, 33-38. http://doi.org/10.1016/j.procs.2014.06.006

[9] Kobos, M., Bolikowski, Ł., Horst, M., Manghi, P., Manola, N., & Schirrwagen, J. (2014). Information Inference in Scholarly Communication Infrastructures: The OpenAIREplus Project Experience. *Procedia Computer Science*, 38, 92-99. http://doi.org/10.1016/j.procs.2014.10.016

[10] Jisc Blog, "Jisc Publications Router enters a new phase".

## About the Authors

**Michele Artini** is a research fellow at the Networked Multimedia Information Systems laboratory of the Istituto di Scienza e Tecnologie dell'Informazione (ISTI), Consiglio Nazionale delle Ricerche, Pisa, Italy. Since 2005 he has been involved in EC funded projects for the realisation of aggregative data infrastructures like DRIVER, DRIVER II, BELIEF, HOPE, EFG, EFG1914, OpenAIRE, OpenAIREPlus and Openaire2020. He is interested in digital libraries, service-oriented infrastructures, database systems and workflow management systems.

**Claudio Atzori** received his MSc in "Information Technology" in 2009 at the University of Cagliari. He is a PhD student in Information Engineering at the Engineering School "Leonardo da Vinci" of the University of Pisa. He works as a research fellow in the InfraScience research group, part of the Multimedia Networked Information System Laboratory (NeMIS), at the "Istituto di Scienza e Tecnologie dell'Informazione", National Research Council, Pisa, Italy. He works on the realisation of aggregative data infrastructures for the e-science and scholarly communication. He has also participated to the EC funded R&D projects: DRIVER-II, EFG, EFG1914, HOPE, EAGLE, OpenAIRE, OpenAIREPlus, OpenAIRE2020.

**Alessia Bardi** received her MSc in Information Technologies in the year 2009 at the University of Pisa, Italy. She is a PhD student in Information Engineering at the Engineering Ph.D. School "Leonardo da Vinci" of the University of Pisa and works as graduate fellow at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. Today she is a member of the InfraScience research group, part of the Multimedia Networked Information System Laboratory (NeMIS). She is involved in EC funded projects for the realisation of aggregative data infrastructures. Her research interests include service-oriented architectures and data infrastructures for e-science and scholarly communication.

**Sandro La Bruzzo** is a research fellow at Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He received his MSc in Information Technologies in the year 2010 at the University of Pisa, Italy. Today he is a member of the InfraScience research group, part of the Multimedia Networked Information System Laboratory (NeMIS). His current research interests are in the areas of Service-Oriented Infrastructures for Digital Libraries, protocols for metadata exchanging, Database, Index. He is currently working for the development of the Digital Library and Data infrastructures for the European Commission projects OpenAIRE, OpenAIREplus, OpenAIRE2020, EFG1914, HOPE, and EAGLE.

**Paolo Manghi** is a Researcher in computer science at Istituto di Scienza e Tecnologie dell'Informazione (ISTI) of Consiglio Nazionale delle Ricerche (CNR), in Pisa, Italy. He is the acting technical manager and researcher for the EU-H2020 infrastructure projects OpenAIRE2020, SoBigData.eu, PARTHENOS, RDA Europe, and EAGLE. He is an active member of a number of Data Citation and Data Publishing Working groups of the Research Data Alliance. In addition, he is an invited member of the advisory boards of the Research Object initiative (Carole Goble, University of Manchester) and of the Europeana Cloud project. His research areas of interest currently are data e-infrastructures for science, scholarly communication infrastructures, and publishing/interlinking of data and experiments, with a focus on technologies supporting open science and digital scholarly communication, i.e. reusing, sharing, assessing all research products, be them articles, datasets or experiments.

**Andrea Mannocci** is a research fellow at the Networked Multimedia Information Systems (NeMIS) laboratory of the Istituto di Scienza e Tecnologie dell'Informazione (ISTI), Consiglio Nazionale delle Ricerche, Pisa, Italy. He is also a PhD student in Information Science Engineering at the University of Pisa. He is involved in EC funded projects for the realization of data integration infrastructures (OpenAIREplus, OpenAIRE2020 and EAGLE). His research activities span data quality and application monitoring, and service-oriented infrastructures for e-science. In 2010 he received a MSc in Computer Science Engineering at the University of Pisa and in 2011 a MSc in Telematic Engineering at the University Carlos III of Madrid, Spain.