RepOSGate: Open Science Gateways for Institutional Repositories

Michele Artini[0000-0002-4406-428X], Leonardo Candela[0000-0002-7279-2727],

Paolo Manghi[0000-0001-7291-3210] , Silvia Giannini[0000-0001-7323-3786]

Institute of Information Science and Technologies
National Research Council, Pisa, Italy
Name.Surname@isti.cnr.it

Abstract. Most repository platforms used to operate Institutional Repositories fail at delivering a complete set of functionalities required by institutions and researchers to fully comply with Open Science publishing practices. This paper presents RepOSGate, a software that implements an overlay application capable of collecting metadata records from a repository and transparently deliver search, statistics, upload of Open Access versions functionalities over an enhanced version of the metadata collection, which include: links to datasets, Open Access versions of the artifacts, links to projects from several funders, subjects, citations, etc. The paper will also present two instantiations of RepOSGate, used to enhance the publication metadata collections of two CNR institutes: Institute of Information Science and Technologies (ISTI) and Institute of Marine Sciences (ISMAR).

Keywords: Institutional Repository, Open Access, Open Science, Scholarly Communication, OpenAIRE

1 Introduction

Open Science [3] publishing principles demand for a scholarly record that (i) is persistently stored into repositories and features all kinds of products, not only scientific literature, (ii) makes use of persistent identifiers for all scholarly entities (e.g. authors, organizations, scientific products, thematic services), (iii) keeps track of the semantic relationships between such objects in the metadata (e.g. citations, supplement material, versions), (iv) keeps an up-to-date record of science evolution, by continuously publishing such links within the metadata of the objects in the repositories, and (v) allows the deposition of multiple versions of a publication, each with its own access rights, to make it clear when a publication is also Open Access. Unfortunately, most institutional repository platforms (e.g. Eprints, DSpace, Invenio) are today unable to fulfill all such requirements at once [1,2,4,6,12]. Old releases, still broadly in use due, simply fail to provide support for PIDs and links, or in some cases make a difference between an Open Access and a Closed version of a publication; more recent releases, which may take these into account, fail instead to keep an

up-to-date linking record as they do not offer APIs to collect updates to the metadata records coming from trusted third-party sources.

This paper presents RepOSGate, a general-purpose software conceived to provide an Open Science view of a repository collection by transparently generating an intersection between the repository metadata collection and other public scholarly communication data sources. RepOSGate fetches the "pivot" metadata collection as exposed by a repository and performs an entity linking procedure based on publication DOIs to enrich such collection with properties and links from other sources: (*i*) the OpenAIRE Research Graph [11] for collecting up-to-date information on publications metadata, and (*ii*) the OpenAIRE's Scholexplorer [5] for collecting up-to-date links between publications and dataset objects. As a result, repository users can access the RepOSGate portal, a gateway that transparently offers discovery and statistic functionalities to an enhanced version of the original repository metadata, including for example abstracts, links to Open Access versions, subjects, bibliographic references, links to datasets, links to software, links to projects, and, when missing, ORCID identifiers of the authors. The Gateway offers also OAI-PMH APIs [9], to expose the enriched metadata collection to third-party consumers.

We will also describe the deployment of RepOSGate to deliver the ISTI Open Portal[1], a gateway developed to promote the scientific publications of the Institute of Information Science and Technologies (ISTI)[2] - an institute of the Italian National Research Council (CNR) - by leveraging access to their open access versions. The ISTI Open Portal offers an Institutional Repository web-based user interface for discovery and statistics on top of an aggregation of multiple sources around the "pivot" collection of ISTI's publication metadata. Another installation of RepOSGate supports a gateway for another CNR's institutes, namely ISMAR, the Institute of Marine Sciences[3].

2. RepOSGate Architecture

RepOSGate has been conceived to make sure that scientists of an institution which already operates an institutional repository whose underlying platform cannot meet Open Science demands, can quickly, and at low cost, meet such demands. For example, the European Commission requires funded researchers to deposit in an Open Access repository, with links to project in the metadata, every article accepted for publication. Many repository platforms offered by institutions to researchers do not meet this basic requirement and researchers end up depositing in open shared platforms, such as Zenodo.org or Figshare. As a consequence, virtuous scientists deposit in two repositories, while others simply deposit once following their most

[1] ISTI Open Portal www.openportal.isti.cnr.it
[2] *Istituto di Scienza e Tecnologie dell'Informazione*, http://isti.cnr.it
[3] *Istituto di Scienze Marine*, http://ismar.cnr.it

urgent obligation. On a different aspect, but with similar drawbacks, such platforms do not leverage the good practice of providing links between datasets and articles, or of providing ORCID identifiers. For institutions this means they cannot support their researchers with the tools to comply with funders mandates and cannot provide their scientists with functionalities to keep their local collection of publications interlinked with the evolving scholarly communication infrastructure.

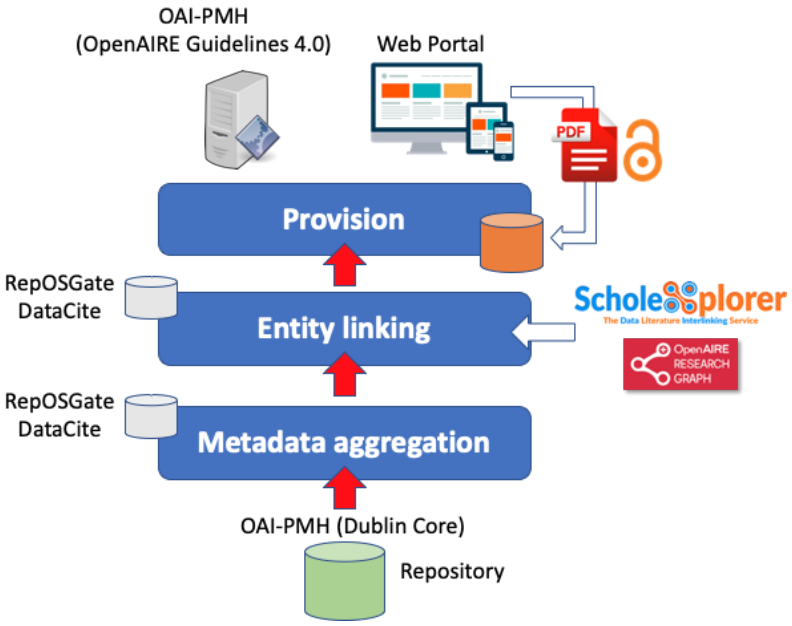

Fig 1. High-level functional architecture

RepOSGate was designed to deliver an overlay platform capable of enhancing content in a repository with up-to-date metadata information regarding their interlinking with projects, datasets, ORCID IDs, Open Access versions, and more. Moreover, repository managers can upload Open Access versions of repository articles via admin interfaces; to facilitate the adoption of RepOSGate, the Open Access files are kept locally to RepOSGate, independently of the repository platform at hand. Ideally, the resulting overlay repository makes the repository OpenAIRE compliant, hence fitting with the OA mandate of the European Commission, and Plan-S compliant[4], since all publications will be equipped with an Open Access version, if it exists.

RepOSGate's software builds on top of a D-NET Toolkit [10] instance. D-NET is a framework where designers and developers can find the tools for constructing and operating so-called aggregative infrastructures, namely systems for aggregating (meta)data sources with heterogeneous data models and technologies. Designers and developers can select from a variety of D-NET data management services, can

[4] *Coalition S*, https://www.coalition-s.org/

configure them to handle data according to given data models, and can construct autonomic workflows to obtain personalized aggregative infrastructures. As shown in Figure 1, RepOSGate adopts D-NET to deliver an aggregation system capable of: aggregation of metadata records from the repository collection, performing entity linking to build richer records, and index the records to expose them via a web portal or via OAI-PMH APIs that are compatible with the OpenAIRE Guidelines 4.0[5].

2.1 Aggregation

The repository must expose the publication metadata records as an OAI-PMH collection of Dublin Core metadata records, where `dc:identifier` should contain the DOI of the record. A D-NET aggregation workflow will be scheduled to harvest the records and transform them into the RepOSGate core metadata schema - the setting up of a D-NET workflow is described in detail in [10] and is not in the scope of this paper. The transformation includes standard harmonization rules to convert country codes, dates, DOI URLs/codes, author names, into a common representation; they can be fine-tuned to match peculiarities of the given repository, for example to include new `dc:subject` or `dc:resourcetype` terms into the vocabulary provided by RepOSGate.[6]

2.2 Entity linking

The entity linking process is based on publications DOIs. The basic methodology is to send requests to external metadata source APIs so as to collect information required to enrich the records. Specifically, RepOSGate has been customized to collect information from three main sources:

- *OpenAIRE Research Graph*: entity linking collects abstracts, links to projects from 28 funders (including MIUR, the European Commission, NSF, Wellcome Trust, and others world-wide), links to other versions of the publication into other sources (possibly Open Access), ORCID identifiers, subjects according to standard vocabularies (e.g. MeSH, DEWEY, Arxiv, ACM, etc.), list of citations in the bibliography;
- *OpenAIRE Scholexplorer*: entity linking collects links from the publication to any dataset referring to it.

The degree of potential enrichment of the "pivot" collection is remarkable considering that:

- The OpenAIRE Research Graph aggregates today, November 2019, around 450Mi metadata records with links, which after deduplication and fine-grained

[5] OpenAIRE Guidelines for Content Providers, http://guidelines.openaire.eu

[6] Note that RepOSGate' vocabularies are the ones of OpenAIRE, which today has a complete coverage of transformation rules for more than 10,000 data sources world-wide.

classification narrow down to ~100Mi publications [11], ~8Mi datasets, ~200K software research products, 8Mi other scientific products, with 480Mi semantic links between them. Such products are in turn linked to 7 research communities, organizations, and projects/grants from ~30 funders worldwide. The Graph aggregates sources such as CrossRef[7], DataCite[8], Microsoft Research Graph[9], Unpaywall, thematic repositories (e.g. ArXiv, RePEc, UK PMC, etc.), all known publishers, journals, data centers, research software repositories, research infrastructure archives/repositories, and all known registries (e.g. ORCID[10], GRID.ac, re3data.org, OpenDOAR[11], etc.). The graph is refreshed with new content every two weeks.

- The OpenAIRE Scholexplorer aggregates article-dataset and dataset-dataset links from publishers and data centers world wide, for a total of 480Mi links (a dump of Scholexplorer is available at [8]); its APIs are used by Scopus and tens of data centres and publishers to resolve DOIs to the relative linked objects. The Scholexplorer citation graphs is being kept refreshed every hour, with sync actions with DataCite and CrossRef EventData.

Each record in the repository with a DOI is enriched by the knowledge stored in the sources above, to build a richer record with up-to-date information.

2.3 Provision

The final step of data provision is that of indexing the enriched records and deliver an OAI-PMH API and Full-Text Index API with a web portal. This is performed by integrating in the D-NET workflow of aggregation and entity linking a final step of ingestion into the D-NET services designed for this specific purpose, namely the OAI-PMH Publisher (based on MongoDB[12]) and the Index Service (based on Apache Solr[13]). The RepOSGate portal is a general purpose UI, which can be configured to include custom branding and text in static pages, which offers search and browse functionalities and statistics on Open Access and Open Science, including integration with Altmetrics to show social media citations to the article DOIs. The user interface allows the upload of Open Access versions of the original PDFs.
The following section will showcase the portal as deployed for the CNR institutes ISTI and ISMAR, whose publication collection is available via People, the central institutional archive of CNR.

---

[7] CrossRef, http://crossref.org
[8] DataCite, http://datacite.org
[9] Microsoft Academic, http://www.microsoft.com/en-us/research/project/academic/
[10] ORCID Researcher Identifiers, http://orcid.org
[11] OpenDOAR Repository Registry, http://opendoar.org
[12] *mongoDB*, https://www.mongodb.com/
[13] *Apache Solr,* https://lucene.apache.org/solr/

3. ISTI Open Portal

ISTI is the Institute of Information Science and Technologies of the National Research Council of Italy, counting around 250 members of staff. CNR researchers are mandatorily requested to yearly report their scientific publications, by depositing metadata and files into the central CNR archive called People. People stores and exposes bibliographic metadata according to an internal qualified Dublin Core, where: (*i*) CNR author enroll numbers are kept with author CNR names, in turn kept separated from non-CNR author names (strings); (*ii*) grant names are kept in special fields as strings, with no reference codes; (iii) for each article the system can acquire multiple files, but no access rights are provided. People is only used for CNR and National research assessment programs, hence does not offer public search, browse, statistic portals nor public OAI-PMH APIs. This means that, unless the CNR will establish and embrace a roadmap to upgrade the current system, the institutes will not be capable of implementing Open Access and Open Science practices at the level required today by research organizations.

Luckily enough, CNR's People APIs are available on request. In order to offer a traditional institutional Open Access repository portal and OAI-PMH APIs, we have deployed an instance of RepOSGate to deliver the ISTI Open Portal[14]. A twin installation[15] has been made available for the ISMAR institute, the Institute of Marine Sciences. In the following we shall present the aggregation and entity linking workflow implemented by RepOSGate for ISTI but also show the numbers for ISMAR Open Portal, to demonstrate the gain in information enrichment.

3.1 Aggregation

RepOSGate collects from People OAI-PMH APIs only the metadata of publications provided by ISTI researchers, via a dedicated OAI-PMH Set. The transformation makes sure that:

- *CNR authors:* CNR Author information, which is properly structured, is included into the DataCite author metadata in such a way CNR enrollment number appears as author identifier;
- *non-CNR authors*: non-CNR Author information follows the same restructuring, but applying a case-driven function that attempts to transform the name into an "Surname, N." structure.
- *ISTI Laboratories*: Thanks to a custom author-laboratory map, CNR authors are also associated to their ISTI Laboratory, the information being kept into the affiliation field of the author structure.

[14] ISTI Open Portal http://openportal.isti.cnr.it
[15] ISMAR Open Portal http://openportal.ismar.cnr.it

Records from People are not clear in terms of Access Rights. This information is key to deliver an Open Access repository or view over the scientific production of ISTI and will be identified via the Entity Linking below.

3.2 Entity Linking

The entity linking process fetches from OpenAIRE Research Graph and Scholexplorer: links to projects, links to other versions of the publication into other sources (possibly Open Access), links from the publication to any dataset referring to it, bibliographic references, and subjects according to standard vocabularies (enrichment with ORCID IDs is in the roadmap).

More specifically, by the 22nd of September 2019 the system collected 9329 publication records, out of which 2872 have DOIs (the majority of publications does not necessarily bear a DOI, for example technical reports, presentations, software, etc.). The administrator has uploaded 360 Open Access versions of non-OA articles. The entity linking phase enriched a total of 590 records by querying the OpenAIRE services, the numbers shown in Table 1. Of all information enrichments above, of great interest to the Open Access and Open Science mission of ISTI is in particular:

- The number of Open Access publications: such numbers could not be identified from the records in People and they are key to offer Open Access analysis and monitoring.
- Identification of Open Access rights: as long as an Open Access version of a non-Open Access publication in ISTI Open Portal will be collected by OpenAIRE, this version will also appear in the ISTI Open Portal as part of the publication metadata: researchers can freely deposit in EC compliant repositories like Zenodo.org to comply to the EC Open Access mandates and this version will be first collected by OpenAIRE and then fetched by ISTI Open Portal;
- Identification of links to funding: for the same reason, the projects funding the publication will be fetched from OpenAIRE by the ISTI Open Portal and will appear as part of the publication record.

Table 1. ISTI Open Portal: metadata enrichment statistics.

Properties and links	Original metadata	Enriched metadata	Difference
Articles with Open Access version	0	757 (360 added by administrator by depositing OA files)	757
Subjects	10752	12744	1992
Bibliographic references	0	6306	6306
Other versions	8146	9270	1124
Project links	610	770	160
Dataset links	0	54	54

Table 2 shows the numbers for ISMAR Open Portal. The 26th of September 2019 the system collected 5891 records, out of which 1898 have a DOI. Interestingly, the marine context features a richer set of datasets if compared to the Computer Science.

Table 2. ISMAR Open Portal: metadata enrichment statistics.

Properties and links	Original metadata	Enriched metadata	Difference
Articles with Open Access version	0	417 (no OA file deposited by administrator)	417
Subjects	7090	10454	3364
Bibliographic references	0	10038	10038
Other versions	4770	6068	1298
Project links	45	189	144
Dataset links	0	128	128

For both ISTI Open Portal and ISMAR Open Portal the aggregation and entity linking process is scheduled every night, thereby keeping the ISTI collection always up to date with the latest scholarly links and properties collected by OpenAIRE services.

3.3 Provision

RepOSGate's web portal offers a number of services including: (a) a per-publication page offering augmented information with respect to that natively stored in the institutional archive; (b) browsing options taking into account the ISTI authors and the research laboratories they belong; (c) a rich array of statistics including scholarly production indices, open access indices, and visits and downloads. It is worth highlighting that by aggregating content coming from several sources the portal is also conceived to provide its managers/curators with statistics and indicators on both information completeness and consistency to use to improve what's natively stored in the CNR archive as well as in the rest of providers. Static pages have been added to provide links to the Institute Open Access policy and curators of the site. The envisaged solution is suitable for any CNR institute as well as for any institution/community willing to build a repository by augmenting the content of its native repository(ies)/archive(s).

Figure 2 shows the home search page and the result list page with details on multiple versions of the article, access rights for each version, best access right for the article (following the ordering: Open > Restricted > Embargo > Closed), Altmetrics numbers, and links to projects in OpenAIRE.

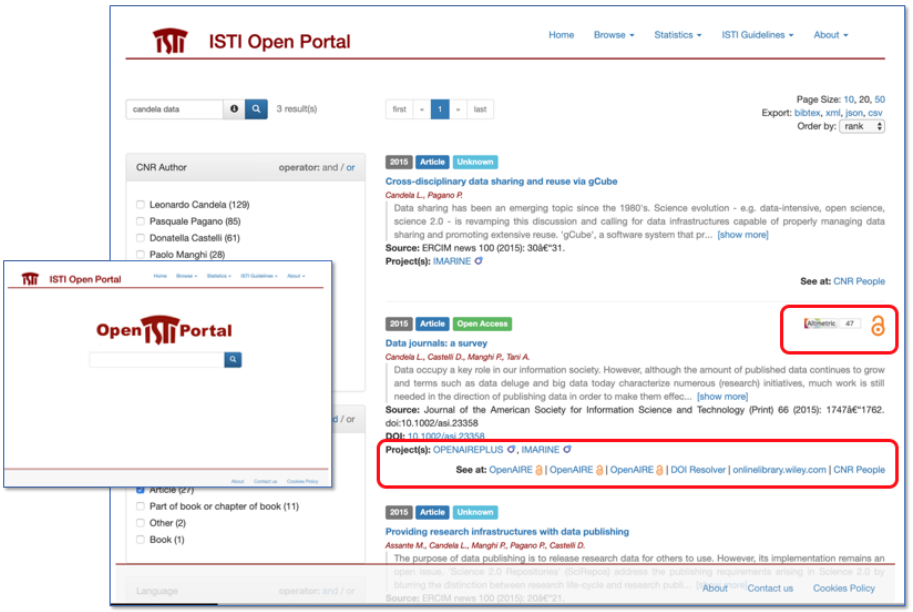

Fig. 2. ISTI Open Portal: home page and search result list.

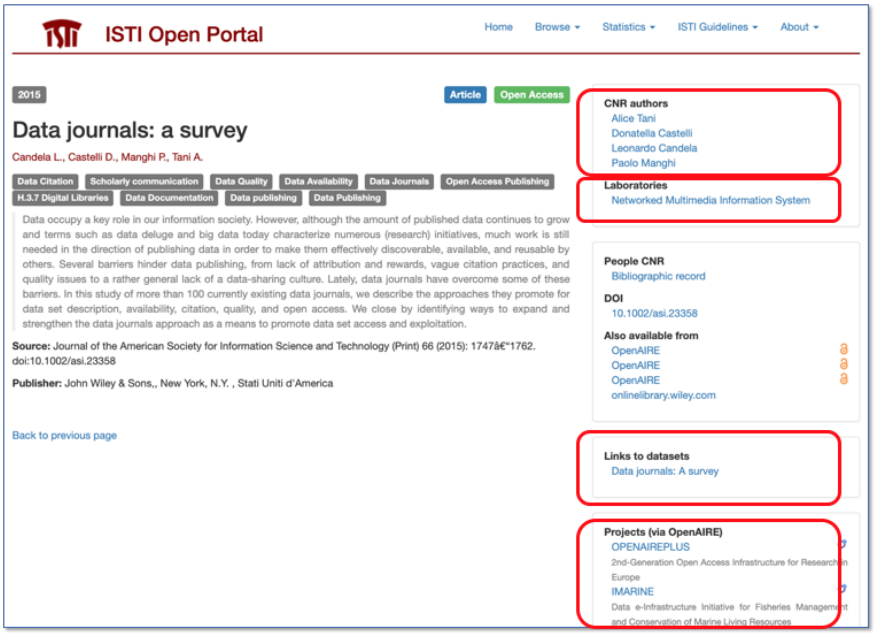

Fig. 3. ISTI Open Portal result details page: author identifiers, ISTI laboratories, and links to datasets.

Figure 3 shows the detail page of the publication “Data Journals: a survey”. This record, which originally featured only the minimal metadata available from the People archive, includes today the link to the related ISTI laboratory, the link to ISTI authors, one DOI link to a dataset returned by OpenAIRE Scholexplorer, and the EC projects funding this work, with links to the detail project pages on the OpenAIRE web site. Figure 4 shows the statistics about the scientific production of ISTI over time, by access rights and by year, both in graph and tabular forms. Other statistics, by year/typology and by laboratory/typology, are shown in Figure 5.

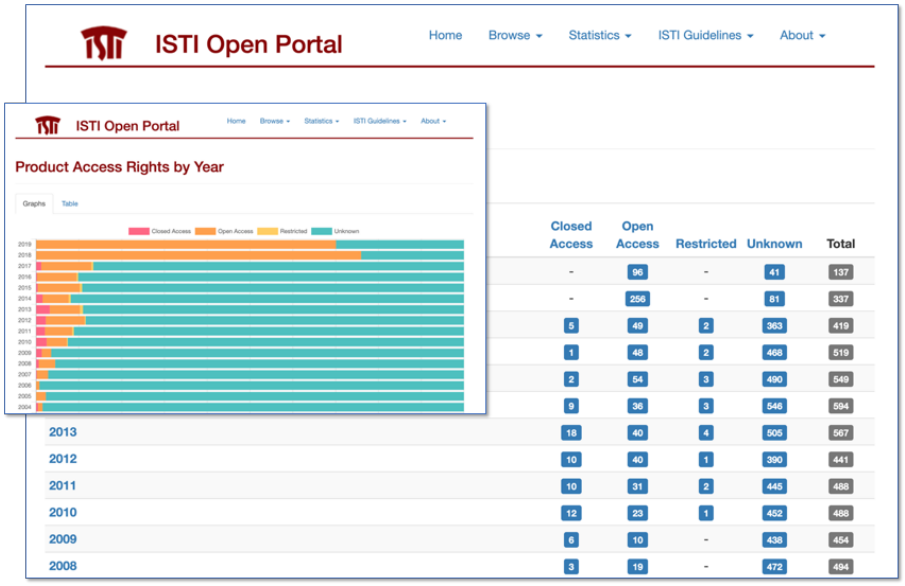

Fig. 4. ISTI Open Portal Open Access statistics by access rights/year: graph and table view.

The increase in open access versions in 2018 and 2019 is primarily due to the emission of the ISTI Open Access Policy since January 2018. From the date of entry into force of the Policy, each ISTI Author must deposit the metadata and a digital version for open access purposes in the CNR institutional archive. The import of the previous digital versions and their access rights is in progress. It should sharply limit the number of "Unknown rights."

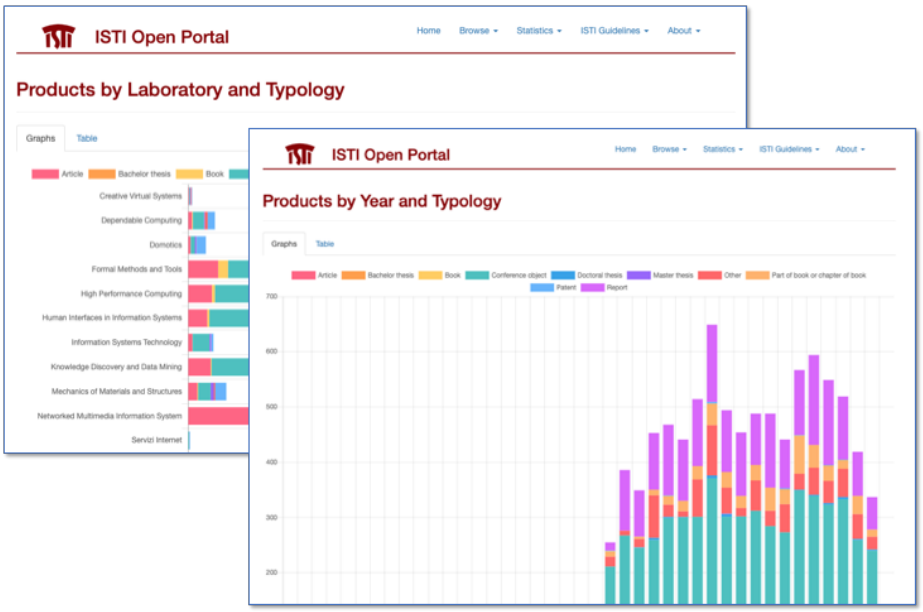

Fig. 5. ISTI Open Portal Open Access statistics by Laboratory/Year and Year/Typology of publications.

4 Conclusion and Prospects

RepOSGate has been developed to provide repository managers with a lightweight solutions easiying the development of their repository with respect to open science practices. This solution benefits from the large amount of knowledge that exists in the scholarly communication web to augment the information accompanying every repository artifact.
The adoption of this solution was instrumental for ISTI to develop and implement an Open Access policy. From 2018 on (the year the open access policy was signed) the Institute managed to make available more than 70% of its scholarly production.
The adoption of RepOSGate is currently being taken into consideration by other CNR institutes. Several enhancements are in the roadmap, such as exploiting entity linking to collect ORCID IDs and, most importantly, the possibility for authorized researchers to upload the Open Access version of an article rather than delegating one administrator of all the work.

Acknowledgments This work was possible and co-funded by the EC H2020 project OpenAIRE-Advance (grant agreement 777541).

References

1. Arlitsch, K. & Grant, C. (2018) Why So Many Repositories? Examining the Limitations and Possibilities of the Institutional Repositories Landscape, Journal of Library Administration, 58:3, 264-281, DOI: 10.1080/01930826.2018.1436778
2. Assante, M., Candela, L., Castelli, D., Manghi, P., Pagano, P. (2015) Science 2.0 Repositories: Time for a Change in Scholarly Communication. D-Lib Mag. 21(1/2) 10.1045/january2015-assante
3. Bartling, S., Friesike, S. (Eds.) (2014) Opening Science - The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer doi: 10.1007/978-3-319-00026-8
4. Bashir, Asma; Mir, Aasif Ahmad; and Sofi, Dr.Zahoor Ahmad, (2019) Global Landscape of Open Access Repositories. Library Philosophy and Practice (e-journal). 2445. https://digitalcommons.unl.edu/libphilprac/2445
5. Burton, A., Aryani, A., Koers, H., Manghi, P., La Bruzzo, S., Stocker, M., Diepenbroek, M., Schindler, U., Fenner, M. (2017) The Scholix Framework for Interoperability in Data-Literature Information Exchange. D-Lib Mag. 23(1/2), https://doi.org/10.1045/january2017-burton
6. COAR Next Generation Repositories Working Group (2017) Next Generation Repositories - Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group
7. La Bruzzo S., Manghi P., Mannocci A. (2019) OpenAIRE's DOIBoost - Boosting CrossRef for Research. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science, vol 988. Springer, doi:10.1007/978-3-030-11226-4_11
8. La Bruzzo, S. & Manghi, P. (2019). OpenAIRE ScholeXplorer Service: Scholix JSON Dump [Dataset]. Zenodo. http://doi.org/10.5281/zenodo.2674330
9. Lagoze, C. and Van de Sompel, H. (2003), "The making of the Open Archives Initiative Protocol for Metadata Harvesting", Library Hi Tech, Vol. 21 No. 2, pp. 118-128. https://doi.org/10.1108/07378830310479776
10. Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D. and Pagano, P. (2014), "The D-NET software toolkit", Program: electronic library and information systems, Vol. 48 No. 4, pp. 322-354. https://doi.org/10.1108/PROG-08-2013-0045
11. OpenAIRE Research Graph, http://catalogue.openaire.eu/service/openaire.openaire_graph
12. Repository Platforms for Research Data Interest Group of the Research Data Alliance (2016): Matrix of use cases and functional requirements for research data repository platforms. DOI: 10.15497/rda00033