41st IAMSLIC Conference "Blue Growth: Motivating Innovations in Acquatic Information Management". Rome, Italy, September 7-11 2015, in conjunction with the 16th EURASLIC Biennial Meeting.

Title: The OpenAIRE Initiative: fostering Open Science for European researchers

Authors: Alessia Bardi, Donatella Castelli, Paolo Manghi
Addresses of Authors: alessia.bardi@isti.cnr.it, donatella.castelli@isti.cnr.it, paolo.manghi@isti.cnr.it

Abstract

The mission of the OpenAIRE initiative is to foster an Open Science e-Infrastructure that links people, ideas and resources for the free flow, access, sharing, and re-use of research outcomes, services and processes for the advancement of research and the dissemination of scientific knowledge. Its scope goes beyond scientific articles recognising that to achieve its full potential, hence enable reproducibility and repeatability of scientific process, scholarly communication should ensure access to the whole plethora of digital products generated by such process, such as research data, software and models. This paper describes the sequence of enhancements applied over the years to the OpenAIRE infrastructure in order to support this vision of scholarly communication. OpenAIRE offers services to collect information about publication, dataset, and software research products from authoritative data sources (e.g. publication/data repositories, CRIS systems) and reconstruct by mining the semantic links between them, enabling the reconstruction of a research context.

Keywords

Scholarly communication, Open Access, Open Science, e-infrastructure, Research life cycle

Body of Paper

The OpenAIRE initiative is the point of reference for Open Access in Europe. Its mission is to foster an Open Science e-Infrastructure that links people, ideas and resources for the free flow, access, sharing, and re-use of research outcomes, services and processes for the advancement of research and the dissemination of scientific knowledge.

OpenAIRE operates an open, participatory, service-oriented infrastructure that supports:

- The realization of a pan-European network for the definition, promotion and implementation of shared interoperability guidelines and best practices for managing, sharing, re-using, and preserving research outcomes of different typologies;
- The promotion of Open Science policies and practices at all stages of the research life-cycle and across research communities belonging to different application domains and geographical areas;
- The development and operation of a technical infrastructure supporting services for the discovery of and access to research outcomes via a

centralized entry point, where research outcomes are enriched with contextual information via links to objects relevant to the research life-cycle;
- The provision of measurements of the impact of Open Science and the return of investment of national and international funding agencies.

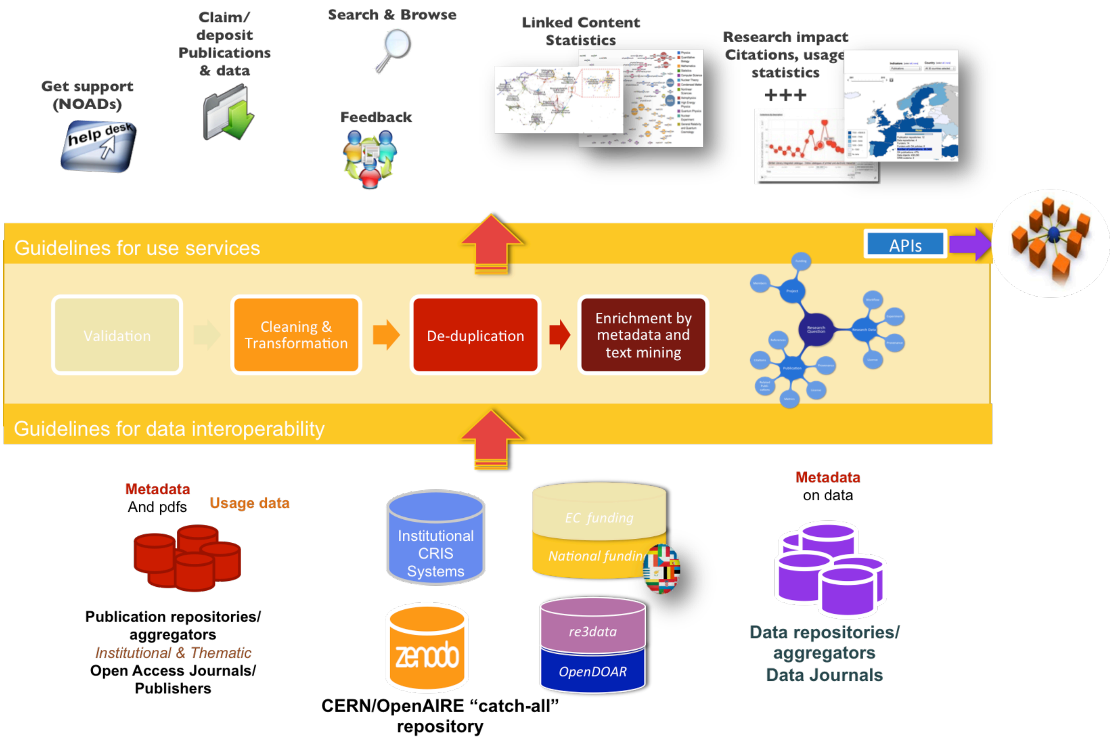
Get support (NOADs)
help desk
Claim/ deposit Publications & data
Search & Browse
Feedback
Linked Content Statistics
Research impact Citations, usage statistics +++
Guidelines for use services
APIs
Validation
Cleaning & Transformation
De-duplication
Enrichment by metadata and text mining
Guidelines for data interoperability
Metadata And pdfs
Usage data
Publication repositories/ aggregators
Institutional & Thematic
Open Access Journals/ Publishers
Institutional CRIS Systems
EC funding
National funding
zenodo
re3data
OpenDOAR
CERN/OpenAIRE "catch-all" repository
Metadata on data
Data repositories/ aggregators
Data Journals

Figure 1 The OpenAIRE infrastructure

The OpenAIRE technical infrastructure (Figure 1) collects information about objects of the research life-cycle from different types of data sources: (i) article metadata and full-texts are collected from institutional and thematic repositories, Open Access journals and publishers; (ii) dataset metadata are collected from data repositories and data journals; (iii) metadata about data sources, organizations, projects, and funding programs are collected from entity registries, i.e. authoritative sources such as CORDA for FP7 and H2020 projects, OpenDOAR for publication repositories, DOAJ for Open Access journals.; (iv) metadata about publications, datasets, persons, organisations, projects, funding, equipment and services are collected through CRIS systems (Common Research Information System) (Houssos et al. 2014, Houssos et al. 2015).

To support the implementation of interoperability guidelines, the OpenAIRE technical infrastructure provides repository managers with a Validation Service, which verifies if metadata records are exported according to the guidelines and, if not, suggests corrections. Infrastructure services collect, harmonize and enrich metadata records compliant to the guidelines, to finally create a graph where objects of the research life cycle are contextualised thanks to semantic relationships. Relationships between objects are collected from the data sources, but also automatically detected by inference algorithms (Kobos et al. 2014) and added by users, who can insert links between publications, datasets and projects via the claiming procedure available from the OpenAIRE web portal (www.openaire.eu).

These “objects in context” are available for human and machine consumption via the OpenAIRE web portal and different kinds of APIs (http://api.openaire.eu).
OpenAIRE also features National contact points, the National Open Access Desks (NOADs), whose aim is to support researchers, project managers, funders, and repository administrators at implementing Open Access policies (to comply with National and European Open Access mandates) and research data management plans for the EC H2020 Open Data Pilot (https://www.openaire.eu/ordp/ordp/pilot).

The current OpenAIRE infrastructure is the result of a sequence of enhancements and extensions applied over the years to better support the scholarly communication and to respond to the needs of research communities and funding agencies (Figure 2).

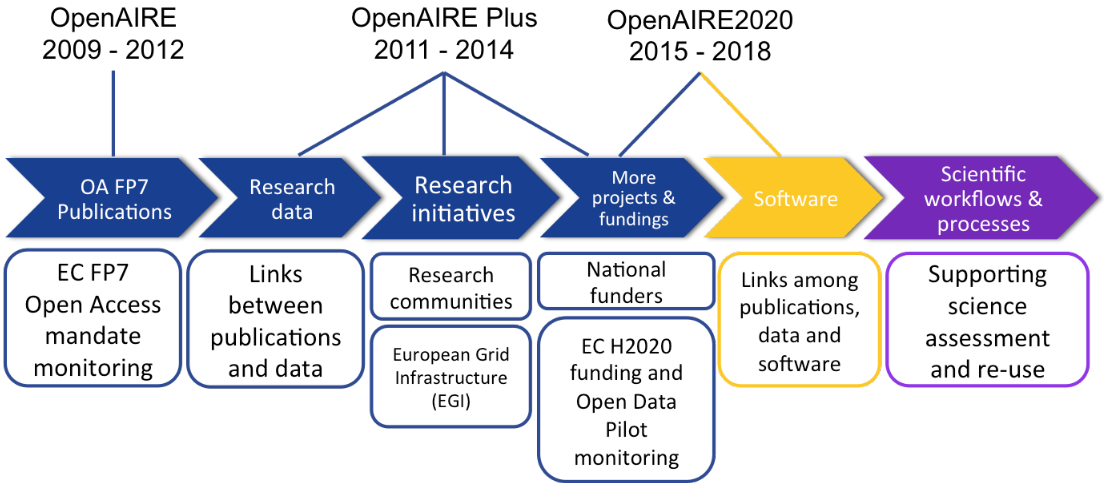
OpenAIRE
2009 - 2012
OpenAIRE Plus
2011 - 2014
OpenAIRE2020
2015 - 2018
OA FP7 Publications
Research data
Research initiatives
More projects & fundings
Software
Scientific workflows & processes
EC FP7 Open Access mandate monitoring
Links between publications and data
Research communities
European Grid Infrastructure (EGI)
National funders
EC H2020 funding and Open Data Pilot monitoring
Links among publications, data and software
Supporting science assessment and re-use

Figure 2 Evolution of the OpenAIRE infrastructure

In its earlier stage, back in 2009, OpenAIRE started, as a pilot project funded by the 7th Framework Program (FP7) of the European Commission (EC), to form a network of interoperable institutional and thematic repositories for the exchange of metadata records about publications deposited according Open Access (OA) policies (Budapest Declaration 2002) and reporting research activities funded by FP7. Thanks to interoperability guidelines for metadata exchange (https://guidelines.openaire.eu), OpenAIRE was able to aggregate metadata records from the network of repositories and monitor the Open Access mandates of the EC (European Commission 2008) and the European Research Council (ERC) (European Research Council 2014).
In 2011 a new EU project, OpenAIREPlus (Manghi et al. 2012a), was funded that enabled OpenAIRE to perform its first evolution step to satisfy requirements of research communities willing to share research data related to publications. New interoperability guidelines were established among the OpenAIRE network to foster the exchange of metadata records about research data. New services were integrated into the infrastructure to enable the deposition of research data (the Zenodo repository: http://www.zenodo.org) and to identify data citation links in the full-texts of publications.

The OpenAIRE network of repositories grew over the years and embraced, together with institutional and thematic repositories, also data repositories (registered in re3data) and OA journals (registered in DOAJ, the Directory of Open Access Journals). In addition, services for the curation and enrichment of aggregated research outcomes were enhanced to detect duplicate publications (which can be possibly collected from different repositories) (Manghi *et al.* 2012b) and to infer descriptive properties (e.g. classification subjects) and links among research outcomes and projects (Kobos *et al.* 2014).

The richness in terms of quantity and quality of research outcomes available from OpenAIRE attracted a number of stakeholders of scholarly communication interested in the analysis of project impact and return of investment of national and international funding agencies. Some examples are the Portuguese Fundação para a Ciência e a Tecnologia, Wellcome Trust and the European Grid Infrastructure (EGI), for which OpenAIRE enabled the computation of statistics to monitor the impact of their projects in terms of OA and non-OA publications.

In 2015, the EC started the new funding program Horizon2020 (H2020) and launched the Open Research Data Pilot, whose goal is to promote best practices on research data management to finally support research reproducibility and effective science assessment (European Commission 2015). The EC endorsed OpenAIRE (now funded through the OpenAIRE2020 project) both for monitoring the pilot and for supporting researchers and project managers in being compliant to the new EC guidelines and best practices.

As of December 2015, OpenAIRE collects from more than 600 data sources, 13 millions (de-duplicated) publications and 10,000 datasets featuring 180,000 links to projects from five different funders: EC FP7 and H2020, Wellcome Trust, Fundação para a Ciência e a Tecnologia, the Australian Research Council, and the Australian National Health and Medical Research Council. Collaborations with Science Foundation Ireland and the government of Croatia are on going and their projects will be integrated in OpenAIRE in the next months.

OpenAIRE is continuously evolving to embrace and advocate new demands of scholarly communication towards achieving Open Science. The next phase will focus on including further typologies of research outcomes, produced or used at different stages of the research life cycle. Since the advent of data-intensive science and the diffusion of Virtual Research Environments (VREs) (Candela et al., 2013) and Research Infrastructures (RIs), software has become a fundamental tool to researchers for carrying out their activities (e.g. the acquisition and elaboration of research datasets, the execution of in-silico experiments and data analysis). As such, scholarly communication workflows should include software as a first citizen, hence foster its publishing, assessment, and re-use. Software publishing though, is just one of the milestones on the road towards Open Science. In fact, Open Science fosters reproducibility of scientific process, which can only be supported when all research products and tools used and produced during research investigations are made available according to agreed on polices and practices. Still, current scholarly communication practices are far from these objectives. A few digital products are entitled as first citizens of scholarly communication (e.g. research data in several sciences, models for some sciences), and best practices and workflows for publishing are generally missing. On top of that, current publishing workflows (Figure 3), conceptually and de-facto separate the place where research is conducted, the RIs, and the place where research products are assessed, published, and shared (marketplace services such as journals and conferences) (Assante et al. 2015). This separation causes products generated in the RI to be published as product copies (snapshots) on marketplace services, these products being rarely described or interlinked with each other like they were in the RI (e.g. links between papers and datasets). Published products therefore lose their context of generation, the nature of their links being missing or doomed to degenerate over time, their original version in the RI possibly evolving to more mature or up to date versions and therefore diverting from their published copies. As a consequence, reproducibility of science is generally mined at its foundations and a lot of work is ahead of us to ensure effective Open Science principles are achieved.

Some initiatives, in line with the OpenAIRE vision, propose to overcome these issues by integrating in the RI functionality for the assessment and publishing of research results. Science 2.0 Repositories (SciRepo) (Assante et al. 2015) represent one of those initiatives and proposes to evolve RIs to support marketplace-like functionality for the publishing and sharing of research products in context (together with related products), in place (where the research has been conducted), and during the research activities (via automatic notification of peers through a social research network). SciRepos endow users of an RI with additional functionality for science reproducibility and assessment, such as:

- Promotion of continuous, in-context, and open peer review process of all research products;
- Support for the production of "altmetrics" to measure the impact of research activities and published products.

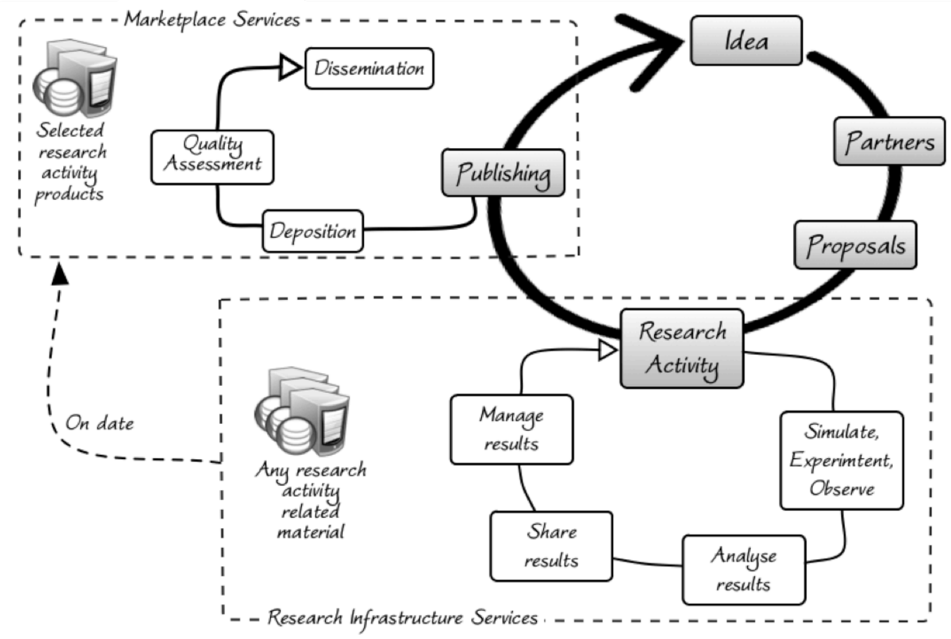

Figure 3 Research lifecycle and publishing of research products (Assante et al. 2015)

However, only a few RIs are today supporting similar tools and in general most scientific disciplines still suffer from the separation of research and publishing environments. OpenAIRE compensate these issues by re-constructing the context of a research a posteriori, collecting products where they are published and identifying relationships among them and with other relevant entities of the research life cycle (e.g. funders, projects, institutions, RIs).

Conclusions

The mission of the OpenAIRE initiative is to contribute to the advancement of research by means of interlinking and disseminating scientific knowledge. Thanks to its growing participatory network (that includes different scholarly communication stakeholders, including institutional repositories, data repositories, OA journals, libraries, and funders), OpenAIRE has a unique opportunity to operate a European entry point to OA scientific knowledge. Thanks to its e-infrastructure services, the European (and beyond) pool of publications, research data, software, workflows, scientific processes, and other research products are aggregated, interlinked, and contextualized to be made easily and openly accessible worldwide. Future plans keep an eye on reproducibility of science and aim at (i) increasing the amount, typology (e.g. software and patents) and linking of accessible research products, and (ii) equipping the infrastructure with other end-user services to further expand the possibilities of end-users and third-party services to access and consume the OpenAIRE graph.

Acknowledgments. Research partially supported by the EC FP7 project iMarine (Grant agreement: 283644, Call: FP7-INFRASTRUCTURES-2011-2), and the EC H2020 projects: OpenAIRE2020 (Grant agreement: 643410, Call: H2020-EINFRA-2014-1) and BlueBridge (Grant agreement: 675680, Call: H2020-EINFRA-2015-1).

References

Assante, M., Candela, L., Castelli, D., Manghi, P., & Pagano, P. (2015). Science 2.0 Repositories: Time for a Change in Scholarly Communication. D-Lib Magazine, 21(1), 4. doi:10.1045/january2015-assante

Candela, L.; Castelli, D., Pagano, P. Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal, Volume 12, July 2013, pp GRDI75-GRDI81 doi: 10.2481/dsj.GRDI-013

Budapest Declaration (2002). Budapest Open Access Initiative, http://www.budapestopenaccessinitiative.org.

European Commission (2008). Annex 1 - Special clauses (clause 39 on Open Access). http://ec.europa.eu/research/press/2008/pdf/annex_1_new_clauses.pdf

European Commission (2015). Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

European Research Council (2014). Open Access Guidelines for research results funded by ERC – revised December 2014.

https://erc.europa.eu/sites/default/files/document/file/ERC_Open_Access_Guidelines-revised_2014.pdf

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. doi:10.1016/j.procs.2014.06.006

Houssos, N., Jörg, B., Dvořák, J. (2015). OpenAIRE Guidelines for CRIS Managers 1.0. Zenodo. doi:10.5281/zenodo.17065

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. doi:10.1016/j.procs.2014.10.016

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

Manghi, P., Mikulicic, M., & Atzori, C. (2012). De-duplication of aggregation authority files. *International Journal of Metadata, Semantics and Ontologies*, *7*(2), 114-130.