An Infrastructure for Managing EC Funded Research Output

– The OpenAIRE Project* –

Paolo Manghi1, Natalia Manola2, Wolfram Horstmann3, Dale Peters4

1 Istituto di Scienza e Tecnologie dell’Informazione “Alessandro Faedo”
Consiglio Nazionale delle Ricerche, Pisa, Italy
paolo.manghi@isti.cnr.it
2 National Kapodistrian University of Athens,
Department of Informatics and Telecommunications, Athens, Greece
natalia@di.uoa.gr
3 University of Bielefeld
Bielefeld, Germany
wolfram.horstmann@uni-bielefeld.de
4 University of KwaZulu-Natal
South Africa
petersd@ukzn.ac.za

Abstract. The OpenAIRE project aims at supporting the diffusion and adoption of the European Commission Open Access mandates among researchers in Europe. It will deliver an organizational/technological infrastructure “supporting mechanisms for the identification, deposition, access, and monitoring of FP7 and ERC funded articles”. The organizational infrastructure will provide a European Helpdesk jointly operated by a European network of National correspondents. The technological infrastructure will be based on state-of-the-art software services of the D-NET Software Toolkit developed within the DRIVER and DRIVER-II projects and the Invenio digital repository software developed at CERN.

1 Scenario and Motivations

“All research builds on former work, and depends on scientists’ possibilities to access and share scientific information. The advent of the internet and electronic publishing, have resulted in unprecedented possibilities for the dissemination and exchange of information. ‘Open Access’, defined as free access over the internet, aims to improve and promote the dissemination of knowledge,

* OpenAIRE is an EC funded project FP7-INFRASTRUCTURES-2009-1 - Grant agreement: 246686
thereby improving the efficiency of scientific discovery and maximising return on investment in R&D by public research funding bodies.”[1]

Although simple in conception, unrestricted availability of research publications and scientific data is still far from reality; the implementation of policies that promote Open Access to these most important products of research has proved to be very challenging. The “Handbook” on “Open Access – Opportunities and Challenges” [2] provides a comprehensive overview of the complexities of all relevant policy, organisational, social and academic cultural, technical, financial, and legal issues. Many initiatives and organisations (SPARC Europe, The Knowledge Exchange1 and The European Universities Association’s expert Working Group on Open Access2) in EU member states and at trans-national level have achieved remarkable progress in the Open Access and Digital Repositories areas over the last years, proving that these complexities can be overcome.

1.1 European Commission Open Access Mandates

In August 2008, the European Commission launched an open access pilot in FP7 [1] under which, grant recipients in seven areas (energy, environment, health, parts of information and communication technologies, research infrastructures, science in society, and social sciences and humanities) will be required to deposit peer reviewed research articles or final manuscripts resulting from their FP7 projects into an online repository, and make their best efforts to ensure open access to these articles within either 6 or 12 months after publication3.

In addition, the ERC Scientific Council’s Statement on Open Access of December 2006 [3] and the subsequent December 2007 Guidelines require that all peer-reviewed publications from ERC-funded research projects be deposited on publication into an appropriate research repository where available, such as PubMed Central, ArXiv or an institutional repository, and subsequently made Open Access within 6 months of publication. ERC also considers essential that primary data are deposited to the

---

1 a cooperation between DEFF, DFG, SURF Foundation and JISC set up in 2005, http://www.knowledge-exchange.info/
3 The pilot covers approximately 20% of the Commission's FP7 research programme budget, and will run until the end of FP7 (2013).
relevant databases as soon as possible, preferably immediately after publication and in any case not later than 6 months after the date of publication.

1.2 OpenAIRE goals and strategy

OpenAIRE (Open Access Infrastructure for Research in Europe) is a three year project that will realize the EC Open Access pilot and assist in the ERC OA guidelines providing an electronic infrastructure and supporting mechanisms for the identification, deposition, access, and monitoring of FP7 and ERC funded articles. Thematically, it will focus on peer-reviewed publications in at least the seven disciplines highlighted in the OA Pilot and on research datasets. Geographically, it will have a definitive “European footprint” by covering the European Union in its entirety, engaging people and scientific repositories in almost all 27 member states.

All deposited articles that result from EU-funded research projects will be freely accessible through the www.openaire.eu portal, which will also provide tools and online support for assisting researchers to publish (deposit) EC funded publications in two ways: (i) by establishing and operating a special repository for articles that can be stored neither in institutional nor in subject-based/thematic repositories, and (ii) by addressing the “depositing only once” issue by developing ways to mediate between the central portal (i.e. OpenAIRE) and the local repositories. Finally, it will provide tools to funders for measuring and evaluating the EC funded research output and the effect of the mandates.

To achieve its goals OpenAIRE concentrates and addresses the following objectives:

**OBJECTIVE 1: Building Support Structures for Researchers in Depositing FP7 Research Publications** - The FP7 OA Pilot mandate has been an indispensable step towards free access to European research results. Experiences with other OA mandates, however, show that acceptance and broad take-up by the scientific community critically depends on accompanying support mechanisms, as any activity beyond the actual research and publishing process is considered by researchers as administrative burden. OpenAIRE will provide support structures assisting them in quick and efficient article depositions, thereby ensuring that a critical mass of articles will be deposited.

**OBJECTIVE 2: Establishment and Operation of the OpenAIRE e-Infrastructure for Peer-
**Reviewed Articles and Other forms of Scientific Results**

OpenAIRE will realize and maintain a system based on state-of-the-art software services of the D-NET software toolkit [8] developed within the DRIVER\(^1\) projects and the Invenio digital repository software developed at CERN\(^2\). These will be further enhanced and complemented with services developed within OpenAIRE to show links of publications to the EC fundec projects. Furthermore, monitoring tools and statistics services will infer relevant information and statistics on FP7 and ERC funded research from article, research data, project metadata and relationships between them.

**OBJECTIVE 3: Exploration of and experimentation with Scientific Data Management Services**

OpenAIRE will work with several subject communities to explore the requirements, practices, incentives, workflows, data models, and technologies to deposit, access, and otherwise manipulate research datasets of various forms in combination with research publications. Feasibility studies will be produced to explore all relevant complex processes & structures and show the benefit for researchers in both depositing and re-using these combined information resource packages. OpenAIRE will investigate supporting potentially all scenarios mentioned earlier about peer-reviewed articles and will choose a subset of those in the actual studies.

**OBJECTIVE 4: Sustainability of the OpenAIRE e-Infrastructure**

Sustainability of OpenAIRE will only be ensured if there are dynamic developments, with new repositories joining the e-Infrastructures, new documents being uploaded, and new researchers making use of the services offered. OpenAIRE aims at marshalling in a user-friendly way, existing strategies to promote the Open Access principle and the use of its e-Infrastructure, and to demonstrate the advantages of integrated efforts with research funding organizations and information infrastructures thus obtained for scientific and scholarly communities.

The project’s support for the establishment of OA offices in all EU member states is particularly important as several European countries have not put Open Access or repositories on their national agendas [4]. This will have significant impact on a broad uptake of Open Access at a pan-European

---

scale, thus ensuring the livelihood of the OpenAIRE tools and services. Moreover, OpenAIRE’s synergies with other relevant initiatives and its close link to COAR\(^1\) - Confederation of Open Access Repositories - will make OpenAIRE’s outcomes visible to a broader, international audience.

1.3 Paper focus and outline

Section 2 provides a synopsis of OpenAIRE’s organizational infrastructure offering services to support researchers over the 27 partner European Countries towards meeting the demands of the FP7 OA mandate (Objective 1). Section 3 presents a more detailed overview of OpenAIRE’s technical infrastructure offering applications for the population, access and manipulation of a European Research Information Space of OA publications and their relationships with the EU administrative project data (Objective 2). Section 4 provides an outline of how OpenAIRE plans to investigate and experiment on data models and functionalities for the future extension of the technical infrastructure with applications for managing research data and how to combine it with related publications in the Information Space (Objective 3). Finally, Section 5 comments on future issues and possibilities.

2 Organizational infrastructure

Authors and researchers in general must be informed on the various aspects of the depositing process related to the OA mandates: they must be aware of issues ranging from different depositing process scenarios (e.g. self or mediated) and guidelines, publishing and licensing models, copyright/licensing policies (e.g. Romeo\(^2\) database on copyright policies), the difference between author versions vs. publisher versions, to policies for requesting authorization to self-archive after embargo. Moreover, researchers need to fulfil the OA terms by operating on the local level, in the context of their affiliation with local universities and research institutes. One of the focal points of OpenAIRE is to provide all the necessary support through a series of tasks carried out in 27 European countries. Its main tasks will be (i) to develop an European Helpdesk System, comprising a European Centre and national Open Access liaison offices, and (ii) to liaise with other Open Access and repository related

---

2 SHERPA RoMEO: publisher copyright policies & self-archiving. [http://www.sherpa.ac.uk/romeo/](http://www.sherpa.ac.uk/romeo/) (Feb 2010)
activities in Europe in order to cascade information to local decision makers in the national context of researcher affiliation to research organizations, or discipline related research activities.

2.1 Pan European Helpdesk System

The provision of a continuous operational service to facilitate deposit is critical to the support envisaged for researchers. The Helpdesk functionality is not limited to standard online query submission and fast response times, but acts as a network providing supporting measures that covers all EU member states, and offers seamless information on issues ranging from the mapping of a point of queries to the location of the nearest relevant repository, localised information on intellectual property rights and deposit guidelines. In particular, the European Helpdesk System will offer:

- FAQ’s, selected list of relevant references (repositories, sites, services, organizations, and experts), as well as relevant materials for incremental distribution to national liaison offices.
- Toolkits for researchers to guide them through the deposition process (advocacy material and information on copyright/licensing policies of selected journals); toolkits for research institutions on OA policy development and compliance with the EC mandates.

The organizational structure of the Helpdesk (Figure 1) is moderated and administered centrally, with distributed responsibility among experienced partners and a network of national OA Liaison Offices.
2.2 Liaison with OA and repository related activities in Europe

The FP7 OA pilot is one of many parts of the developing landscape of Open Access activities in Europe. The overlapping affiliation of OpenAIRE participating organizations with major ongoing initiatives and its strong link to COAR, ensures collaboration with, and inclusion of all relevant stakeholders. The envisioned networked liaison scheme includes building collaborative initiatives with publishers through the PEER project¹, with organizations developing and maintaining Current Research Information Systems (CRIS), and with major European OA organizations (SPARC Europe and the OA Working Group of the European University Association).

3 Technological infrastructure

The OpenAIRE portal (www.openaire.eu) is the gateway to all underlying applications and services provided by the technical infrastructure, where researchers or organizations come to deposit their research outcome, search and access information on related publications or European projects, receive information on OA policies, licenses and copyright issues, or even discover a link to an institutional

¹http://www.peerproject.eu/
or subject repository matching their location or discipline. The targeted user typologies are:

*Authors* interested in depositing their publications into the infrastructure associating it with the FP7 project information;

*Researchers and general public* interested in accessing the Information Space of all OA FP7 publications through friendly web interfaces and a variety of functionalities;

*Commission and organizational funders* studying and assessing the effectiveness of their policies on research directives as well as publication models;

*Repository community* interested in becoming an integral part of the OpenAIRE infrastructure integrating deposition and publication referral processes;

*E-Science applications* operated by third-party organizations, interested in accessing/retrieving content from the Information Space.

Concrete scenarios have been defined for the construction of an integrated European research information system. Although publication deposition and their relation to projects is the central use case, other real-case scenarios, equally important, help at formulating the system components.

**Scenario 1: Depositing a publication in the European Research Information Space**

The system will support two high-level deposition patterns.

*Autonomous deposit:* authors deposit autonomously eligible FP7 and ERC-funded results in existing repositories without direct relation to OpenAIRE.

*OpenAIRE mediated deposit:* Authors come to OpenAIRE and expect guidance for deposition towards existing repositories, or deposit directly into the OpenAIRE provided “orphan repository”

**Scenario 2: Accessing the European Research Information Space**

End-users want to retrieve information on publications and project data by: (i) Searching/browsing the

---

1 Repository for “orphan” researchers, i.e. researchers that do not have an institutional of thematic repository of reference.
content, e.g. breakdown and display by research, projects, people, funds, publications, subjects, etc., (ii) Converting publications on-demand into other types of MIME formats; (iii) Receiving notifications on events related to their expertise and type of work, such as: impact (derived from overall usage statistics) on their publications or projects; new projects or publications appearing within their scientific research area; publications released from specific projects or more general research areas or on usage statistics (calculated by thresholds) on specific research areas/projects.

**Scenario 3: Monitoring and assessing impact through usage statistics**

Users want to measure the impact of publications and projects through usage statistics (obtained from the orphan repository and all other participating repositories) and their combination of various quantitative web indicators involving external link references and public search engines data.

**Scenario 4: Third-party services accessing the Information Space**

Since OpenAIRE is envisioned to be a part of the European e-infrastructure ecology, its content may be re-used by third-party services, e.g. aggregators, distributed search engines, e-Science text-mining applications. Such services and e-infrastructures (e.g. Europeana\(^1\), DRIVER) will be able to retrieve the OpenAIRE rich content through a variety of protocols.

### 3.1 Architecture and technologies

The OpenAIRE technical e-infrastructure provides a system capable of supporting the functionalities identified by the scenarios above, integrated with the tools of the European Helpdesk to present end-users with one, integrated point of entry. As illustrated in Figure 2, the system is structured into four main functional areas: *presentation area and deposition area*, dealing with users accessing and populating the Information Space (system front-end, scenarios 1 and 2); *storage area and mediation area*, dealing with Information Space content management, including storage and access through standard protocols from third party-services (system back-end, scenarios 3 and 4).

---

The system consists of a number of service-oriented applications, “glued” together with the infrastructural application framework offered by the D-NET Software Kit [5],[8], a component oriented approach based on a Service Oriented Architecture (SOA) [6][7], developed during the DRIVER projects. D-NET service applications are open: extendable with new service typologies, i.e. new functionalities; interoperable: services can encapsulate heterogeneous technologies; flexible in architecture: services running on the same site behave like components of a centralized system.

The D-NET Software Kit services are logically organized in three architectural layers:

**Functionality Layer Service Kit.** Includes the services needed to form portals: configurable user interfaces (portals), recommendation systems, user and community profiling, record collections management etc. Appropriately combined, such services can form a variety of community applications, which can then be applied to operate over an information space.

**Data Layer Service Kit.** Includes the services needed to form aggregation system applications, whose functionalities feature harvesting, cleaning, storing, indexing, searching the metadata records harvested from external OAI-PMH Repositories.
Enabling Layer Service Kit. Includes the Services supporting the application framework. These provide “gluing” functionalities such as service registration, discovery, orchestration, user authentication and authorization, subscription and notification.

Figure 3. OpenAIRE services running on D-NET framework.

The choice of D-NET for OpenAIRE was motivated by four main factors: (i) its “gluing features” and “openness”, enabling for technological interoperability and future expansion; (ii) its flexible architecture, supporting both centralized and distributed scenarios; (iii) its proved success in other projects (DRIVER and EFG\(^1\)) for analogous services; (iv) many of the existing services used in DRIVER (harvesting, aggregating, storing, searching) will be adapted and transferred to OpenAIRE.

The design and development of the system is organized in three main steps described in detail in the following sections: Setting up the OpenAIRE Information Space and Orphan Repository, Populating the OpenAIRE Information Space, Accessing and elaborating the OpenAIRE Information Space.

3.1.1 Setting up the OpenAIRE Information Space and Orphan Repository

OpenAIRE supports an Information Space containing publication and administrative EU funded projects data. The system operates under a common data model (the OpenAIRE data model), formally describing the structure of the content: how publication and project metadata are structured and associated with each other, extended to include research data descriptions (data sets) and usage

\(^1\) European Film Gateway: [http://www.europeanfilmgateway.eu](http://www.europeanfilmgateway.eu) (Feb 21010)
statistics data with the corresponding relationships to projects and publications. The Information Space Service offers efficient storage, access and management to the linked OpenAIRE objects, so as to support depositing (Scenario 1) and accessing (Scenarios 2, 3, 4) of publication-project metadata, as well as usage statistics concerning all FP7 OA publications.

![Diagram of OpenAIRE data model]

Figure 4. Indicative example of the OpenAIRE data model.

3.1.2 Populating the OpenAIRE Information Space

The OpenAIRE system enables deposition and ingestion mechanisms (scenario 1) of publication and project metadata. Authors accessing the OpenAIRE portal to deposit their publications are driven through one of the following scenarios:

Manual deposition from authors, where they (i) either deposit the full payload (e.g. pdf, data set) into the OpenAIRE ‘orphan’ repository, or (ii) deposit the “information about the publication” (i.e. bibliographic metadata) while the publication is being, or has been deposited in an OA repository. For (i), the Deposition Service is closely linked to the orphan repository, integrating its UI, user identity management and storage, while for (ii), the Deposition Service provides visual tools enabling authors to identify their publication (e.g. provide repository name, publication DOI), and then to complete the ingestion by linking it to the authoritative EU project metadata provided by OpenAIRE.

Ingestion of publications from repositories, in which case the most important task is to identify and automatically import the publications with the EC-funded footprint, and assign these to the
corresponding projects. Technical solutions require the synergy of OpenAIRE and the repository community in order to define and apply common guidelines for representing and exposing EC-funded publications metadata, i.e. bibliographic and project information, and ultimately implement solutions for notifying OpenAIRE when new publications have been deposited in a repository. Where common guidelines cannot be fully applied, manual and intellectual curation activities verify and fix the operation of this automatic process.

*Ingestion of authoritative EU funded project data from the Commission’s Participant Portal.* The Participant Portal is the gateway to the database maintained by the European Commission to store all its funding related activities (i.e. proposals, projects, institutions, project officers and coordinators, etc.). OpenAIRE, with its extended services, retrieves all the *authoritative* project information and ingests it into its Information Space according to the OpenAIRE data model.

### 3.1.3 Accessing and elaborating the OpenAIRE Information Space

The OpenAIRE system provides interactive access (scenarios 2, 3) to publication and project metadata in the Information Space. Through the OpenAIRE web portal end-users are able to search, browse, access the content. A prototype framework for format conversion services is provided in order to enable users to access documents according to their format of preference.

Through the development of usage statistics services OpenAIRE exploits usage metrics as a supplement to conventional citation analysis (usage metrics are a promising new field of research indicators and citation analysis is too slow to yield results for publications resulting from FP7 and ERC funding within the lifetime of the project). Based on recent or emerging standards and protocols (PIRUS2, SUSHI [9]) different metrics for usage and aggregation activities from various distributed data sources are developed. Usage data will be harvested and aggregated from local institutional or subject-based repositories and the OpenAIRE portal itself in the form of logging usage events (hits, downloads etc.) with server sided tools (e.g. apache Log-files). In combination with ‘webometric’

---

1 The PIRUS2 Project, (JISC), [http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php](http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php) (Feb 2010)
data (“web-citations”, backLinks)\(^1\) and regular citations OpenAIRE intends to provide impact monitoring. Presentation of usage data is planned to take two different forms: (i) an article that is represented through the OpenAIRE portal has additional information on usage in the form of “downloaded n times” (aggregated usage information from local repositories), “accessed through OpenAIRE n times”, “linked from sites a, b, c …” and “cited by x, y, z…”; (ii) comparative reports (written or software based) as a monitoring activity.

Finally, third-party providers may access the information space in a non-interactive way (scenario 4). This involves organizations willing to take advantage of the enriched research information (e.g. links between publication and project data) and its derived results (e.g. usage statistics). Standard protocols, namely OAI-PMH and SRW are delivered, but OpenAIRE also intends to provide customized APIs for the interaction with organizations that are not compliant with the above protocols.

4 Research data and publications

One of the key objectives of OpenAIRE is to perform studies on subject specific requirements for primary and secondary research data in the seven research areas of the FP7 pilot. The expected outcome of such investigation is a book that describes the data models that best capture structure and semantics of the relation between publications and research data in each of the areas to be studied. In particular, the project focuses the analysis on four out of the seven research areas of the Pilot: ‘health’ as exemplified by basic research in the life sciences, ‘environment’ as exemplified by research on climate and agriculture, ICT (Challenge 2: Cognitive Systems, Interaction Robotics), and socio-economic sciences and humanities. The input from those areas is provided by five project partners who are heavily involved in other major European projects:

\(^1\) [http://repositories.webometrics.info/]
Health (Life Sciences): The European Bioinformatics Institute\(^1\) (EBI) as part of the European Molecular Biology Laboratory (EMBL)\(^2\) connecting, among other high-impact services and projects in basic life-science research and health.

Environment: The World Data Center for Climate\(^3\) (Max-Planck Institute for Meteorology\(^4\)) and the Consultative Group on International Agricultural Research (CGIAR\(^5\)) researching the sustainable use and conservation of agricultural biodiversity.

Information & Communication Technology (Cognitive systems, Interaction, Robotics): The Center of Excellence: Cognitive Interaction Technology\(^6\) (CITEC) at Bielefeld University.

Socio-economic Sciences and Humanities: Data Archiving and Networked Services\(^7\) (DANS) connecting to European research infrastructures.

5 Conclusions and Future work

OpenAIRE and the FP7 pilot have started out an important and long term vision for EC funded research. If the pilot will prove to be successful, namely the deposition rate in OA repositories will reach approximately 60%, the OpenAIRE infrastructure will be continued in future EC programmes (FP8), richer content (research data) and greater volumes of publications and statistics will be managed, and other sub-programmes and disciplines, beyond the seven areas of the pilot, will take advantage of the infrastructure (e.g., the Future Emerging Technologies programme in bio-medical sciences).

With respect to research (primary) data deposits, supplementing the text-publication, the challenge is somewhat more complicated. The nature of such materials is so diverse in different subjects, and their manipulation requires such different tools and facilities, that it will become indispensable to provide

\(^1\) [http://www.ebi.ac.uk/](http://www.ebi.ac.uk/)
\(^2\) [http://www.embl.de/](http://www.embl.de/)
\(^3\) [http://www.mad.zmaw.de/](http://www.mad.zmaw.de/)
\(^5\) [www.bioversityinternational.org](http://www.bioversityinternational.org)
\(^7\) [http://www.dans.knaw.nl/en/](http://www.dans.knaw.nl/en/)
subject-specific functionality to be able to cope with the variety of requirements they put to the system, while tailoring to the specific needs of each discipline.

Finally, as the OpenAIRE infrastructure becomes more popular and deposition practices mature, the additional scenario of remote deposition, where researchers use OpenAIRE to deposit publications whose payload is automatically ingested in all authors repositories of reference, will be engineered and relative technical solutions will devised.

Acknowledgments

This work is funded by FP7 Combination of Collaborative projects and Coordination and Support Actions - Integrated Infrastructures Initiative project (I3) proposal, with grant agreement: 246686 (call: FP7-INFRASTRUCTURES-2009-1). This work would have not be possible without the support, ideas and experience of partners and people from the OpenAIRE Consortium.

References


[4] Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU, K. van Eijndhoven, M. van der Graaf http://www.driver-support.eu/documents/DRIVER%20Inventory%20study%202007.pdf


[8] *D-NET Project*: [http://www.d-net.research-infrastructures.eu](http://www.d-net.research-infrastructures.eu), Istituto di Scienza e Tecnologie dell’Informazione, Centro Nazionale delle Ricerche (Pisa, Italy), ICM Research Centre (Warsaw), University of Bielefeld Library (Bielefeld, Germany), and Department of Informatics, National and Kapodistrian University of Athens (Greece)


[http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=2de0e3e04f3a7e32d45db8ee87574c3c8206ddeb](http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=2de0e3e04f3a7e32d45db8ee87574c3c8206ddeb) (Feb 2010)