PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

September/October 2012
Volume 18, Number 9/10

# OpenAIREplus: the European Scholarly Communication Data Infrastructure

Paolo Manghi, Institute of Information Science and Technologies, National Research Council, Pisa, Italy
paolo.manghi@isti.cnr.it

Lukasz Bolikowski, University of Warsaw, Interdisciplinary Centre for Mathematical and Computational Modelling Warsaw, Poland
l.bolikowski@icm.edu.pl

Natalia Manola, National and Kapodistrian University of Athens, Department of Computer Science Athens, Greece
natalia@di.uoa.gr

Jochen Schirrwagen, Bielefeld University Library, Bielefeld, Germany
jochen.schirrwagen@uni-bielefeld.de

Tim Smith, CERN, Geneva, Switzerland
tim.smith@cern.ch

doi:10.1045/september2012-manghi

## Abstract

OpenAIRE and OpenAIREplus (Open Access Infrastructure for Research in Europe) are EC funded projects (Dec 2009 – May 2014) whose goals are to realize, enhance, and operate the Open Access European scholarly communication data infrastructure. This paper describes the high-level architecture and functionalities of that infrastructure, including services designed to collect, interlink and provide access to peer-reviewed and non-peer reviewed publications (from repositories), datasets (from dataset archives), and projects of the European Commission and national funding schemes (from CRIS systems).

## 1. OpenAIRE and OpenAIREplus projects

OpenAIRE (Open Access Infrastructure for Research in Europe) is a three year project (Dec 2009 – Nov 2012) whose goal is to realize and maintain the "Open Access European scholarly communication data infrastructure". The project realizes the European Commission (EC) Open Access pilot[1] and assists in the dissemination and uptake of the European Research Council (ERC) OA guidelines.[2] The aim of the pilot is to ensure application of the OA mandate across Europe and monitor its impact with the support of proper statistics, e.g., OA vs. non-OA peer-reviewed publications per projects. To this aim, OpenAIRE delivers a data infrastructure capable of collecting and monitoring 7th Framework Programme (FP7) and ERC funded research articles across Europe. On the "human" networking level the project operates a European Helpdesk System, comprising a European Centre and National Open Access Desk liaison offices (NOADs), which serve the European Union in its entirety by engaging people and scientific repositories in almost all 27 member states.[3] The NOADs liaise with other Open Access and repository-related activities in Europe (e.g., COAR, SPARC Europe, LIBER) and exploit their hierarchical organization to efficiently disseminate best practices, guidelines, initiatives, and events related to

OA among local decision makers and research organizations (and vice versa). On the technological level, the infrastructure implements services whose aim is to measure the impact of EC OA mandates by collecting and interlinking repository content (i.e., publication metadata) with EC funding information (i.e., CRIS-like metadata about projects, organizations, and people involved). It also provides a so-called "orphan repository" where authors without a repository of reference can deposit their publication files and metadata.

The OpenAIREplus project (Dec 2011 — May 2014) continues and extends the scope of the OpenAIRE data infrastructure to promote and monitor Open Access to a wider audience and to more research output typologies. More specifically, it aims to grow a richer graph of data, covering material from all research disciplines and additional countries, and include projects from national funding schemes, non-peer-reviewed publications, and research datasets. Among its major objectives are experiments to interlink datasets and publications across different disciplines, by automatically inferring semantic relationships between them, by enabling end-users to construct "enhanced publications, and by interoperating with existing infrastructures, e.g., DataCite, Mendeley, ORCID, EUDAT, RElsearch. The OpenAIRE orphan repository scope will also be extended from publications to include datasets.

This paper presents the high-level architecture and functionalities of the data infrastructure services devised in the context of the OpenAIRE project in order to collect and interlink EC project metadata (CRIS-like information) with publications metadata (repository content) in several ways. Furthermore, we describe extensions to the infrastructure envisaged by the OpenAIREplus project, which targets interlinking of both peer-reviewed and non-peer reviewed publications (from repositories), with datasets (from dataset archives), and projects from EC and national funding schemes (from CRIS systems).

## 2. OpenAIRE Data Infrastructure Services

The software services of the OpenAIRE data infrastructure are based on the D-NET Software Toolkit. D-NET offers an open and scalable service-oriented environment where complex data processing workflows can be easily designed, flexibly modified, and extended with new service functionalities. D-NET offers a rich set of data management services, which have been re-used and extended in OpenAIRE to deploy a data infrastructure supporting the following "layered" functionalities: *(i)* management of a dynamic set of heterogeneous data sources, *(ii)* population of an information space by "normalizing" (transformation and cleansing) and interlinking data collected from such sources, and *(iii)* supporting access to the information space to end-users and third-party systems. Figure 1 below shows the D-NET services currently operative in the OpenAIRE infrastructure (the boxes outlined in red indicate services to be realized in OpenAIREplus).

### Management of data sources

OpenAIRE collects and integrates data from various external data sources, which are presented below:

- *Bibliographic metadata records* from institutional repositories which comply with the *OpenAIRE guidelines*; repositories obeying the guidelines must export (through OAI-PMH APIs) Dublin Core records that contain references (project grant agreement numbers) to the EC FP7 projects which funded the publication and a simple access rights specification (open access, embargoed, restricted);
- *The PDF files of publications* from repositories or publishers whose restrictions or bilateral agreement makes this possible; PDFs are used for the sole purpose of text mining to identify FP7 project references;
- *The official list of EC FP7 projects* from the EC CORDA database, which delivers information about project participants;
- *The official list of European institutional repositories* from the OpenDOAR directory.

CORDA and OpenDOAR data sources are integrated in the infrastructure through dedicated synchronization services, which keep the collections up-to-date at given time intervals.

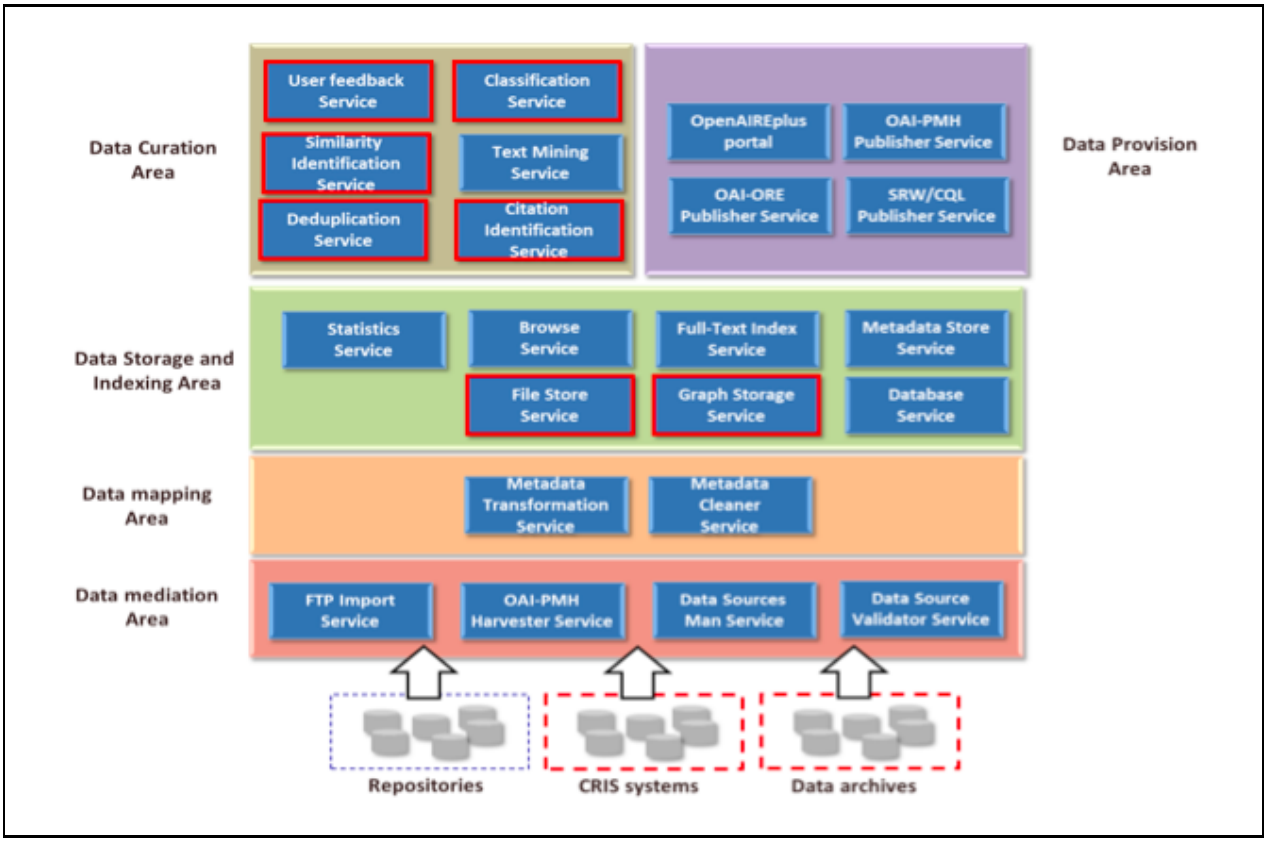


*Figure 1: OpenAIREplus service architecture (red outlined boxes: OpenAIREplus services)*

Repository managers can register their repository to the data infrastructure by following a registration procedure from the OpenAIRE portal, with a pre-requisite that the repository be listed in OpenDOAR. The portal also offers a validation tool through which managers can get feedback on the degree of compliance to the guidelines of their repository. The collection of metadata records from registered repositories is the last step of a workflow consisting of consecutive registration, validation, harvesting, and transformation phases. The whole workflow is automated and supervised by infrastructure data curators.

On the other hand, the acquisition of PDF file collections is negotiated by data curators (on behalf of the Consortium if special non-disclosure agreements are necessary) with the repository managers. Collection and processing of data, which aims at enriching the information space with publication-project relationships, is performed on a case-by-case basis and is described below.

Bibliographic metadata records can also be deposited in the infrastructure directly by end-users who "claim" the related publications through the OpenAIRE portal. This is typically the case for authors whose repository of reference is not yet OpenAIRE compliant. Authors can identify the publications to be claimed by providing the DOIs or by searching through the DRIVER information space, and eventually completing the bibliographic metadata with EC funding information. For those authors who do not have an institutional repository of reference, yet wish to respect the OA mandates of the EC, OpenAIRE operates a repository (based on the INVENIO technology) known as the Orphan repository. The Orphan repository is registered to the infrastructure as an OpenAIRE compliant repository.

**Information Space population**

The OpenAIRE data model encodes a graph consisting of entities, including publications, projects, FP7 schemes, organizations, persons (authors and organization contact points), and data sources (repositories), together with relationships between them. Different data sources map data onto different parts of the data model graph: for example CORDA injects data into the projects-organizations-persons subpart, OpenDOAR into the data sources-organizations part, and repositories into the persons-publications-data sources-projects part. With this aim, the infrastructure processes all incoming data, in order to convert it into the OpenAIRE data

model structure and normalize its semantics accordingly (e.g., vocabulary conversion, format conversion).

**Information Space curation and enrichment**

The infrastructure offers services which automatically infer publications-projects relationships by extracting and mining the PDF text of given publication collections in search of references to FP7 projects. Such services prompt project coordinators with the list of candidate relationships identified for their project, in order to confirm or reject the service guesses. Moreover, in order to provide reliable statistics, the infrastructure copes with duplication of information by offering services capable of disambiguating the information space by identifying and "merging" duplicated publications and authors.[4]

**Access to the Information Space**

Concerning data access, the data infrastructure currently offers a portal which implements advanced search and browsing mechanisms as well as access to statistics. From the same portal, end-users can manage their claims and repository managers can register their repositories to the infrastructure or validate their content with respect to the guidelines. Third-party systems can access all entities of the OpenAIRE information space through dedicated REST, OAI-PMH, OAI-ORE, and SRW/CQL interfaces.

## 3. OpenAIREplus Data Infrastructure Services

The OpenAIREplus project extends the scope of the OpenAIRE data infrastructure services to include and handle in its data model all disciplines covered by EC grants and national funding schemes, and all OA articles, not just published ones, and to absorb the information space of the DRIVER infrastructure, which today includes 6,000,000 bibliographic metadata records relative to Open Access publications from approximately 320 institutional repositories. It will also extend the scope to include research datasets, either via metadata which references the data in OA thematic repositories, or via the Orphan for deposition of datasets without a repository of reference.

By considering specificities in researchers' practice and infrastructures in social sciences, life sciences and atmospheric sciences, link management of text-based publications and datasets in subject-specific infrastructures is investigated. The prototypical exchange and enrichment of information to and from the OpenAIREplus infrastructure will be further exploited in the development of services for the management of "enhanced publications". These are publication-centric compound objects consisting of the textual publication and semantically linked related entities, e.g., associated publications, persons, datasets. As a consequence, the data infrastructure will have to adapt to the new challenges by upgrading its services or introducing and integrating new services in the data processing workflows (see Figure 1).

**Management of data sources**

The infrastructure will have to handle new typologies of data sources, beyond OpenAIRE compliant repositories: Open Access institutional repositories (i.e., delivering metadata about peer-reviewed and non-peer-reviewed Open Access publications without references to projects), CRIS systems (i.e., research management systems delivering CERIF-flavored metadata, namely projects-organizations-publications), and dataset archives (i.e., delivering metadata about datasets, in some cases related to publications).

In order for their metadata to be collected into the infrastructure, distinct data source typologies will have to go through dedicated registration, validation, collection, and transformation workflows. To this aim, OpenAIREplus will draft "guidelines for content providers" which will extend the OpenAIRE guidelines to cover CRIS systems, dataset archives, and institutional repositories which are not OpenAIRE compliant.

The Orphan repository will be enhanced to host research datasets, following the same principles as orphan publications. This will support such datasets to be web published, get DOIs and be connected into enhanced publications. The Orphan repository platform will be extended to support the additional storage and access requirements of research data. This data will be treated as objects without any consideration for the

particular internal schema.

### Information Space population

The OpenAIREplus data model extends OpenAIRE's to include entities such as *results* (e.g., datasets and publications), *licenses* (IPRs for publications and datasets), *funding schemes* (including FP7 schemes), *data sources* (e.g., dataset archives, repositories, CRIS systems) and relationships between them. In order to cope with the encoding of new relationships between such entities, which may vary over time, the data model has been aligned with the CERIF data model *Semantic Layer*.[5] Figure 2 depicts the high-level data model.

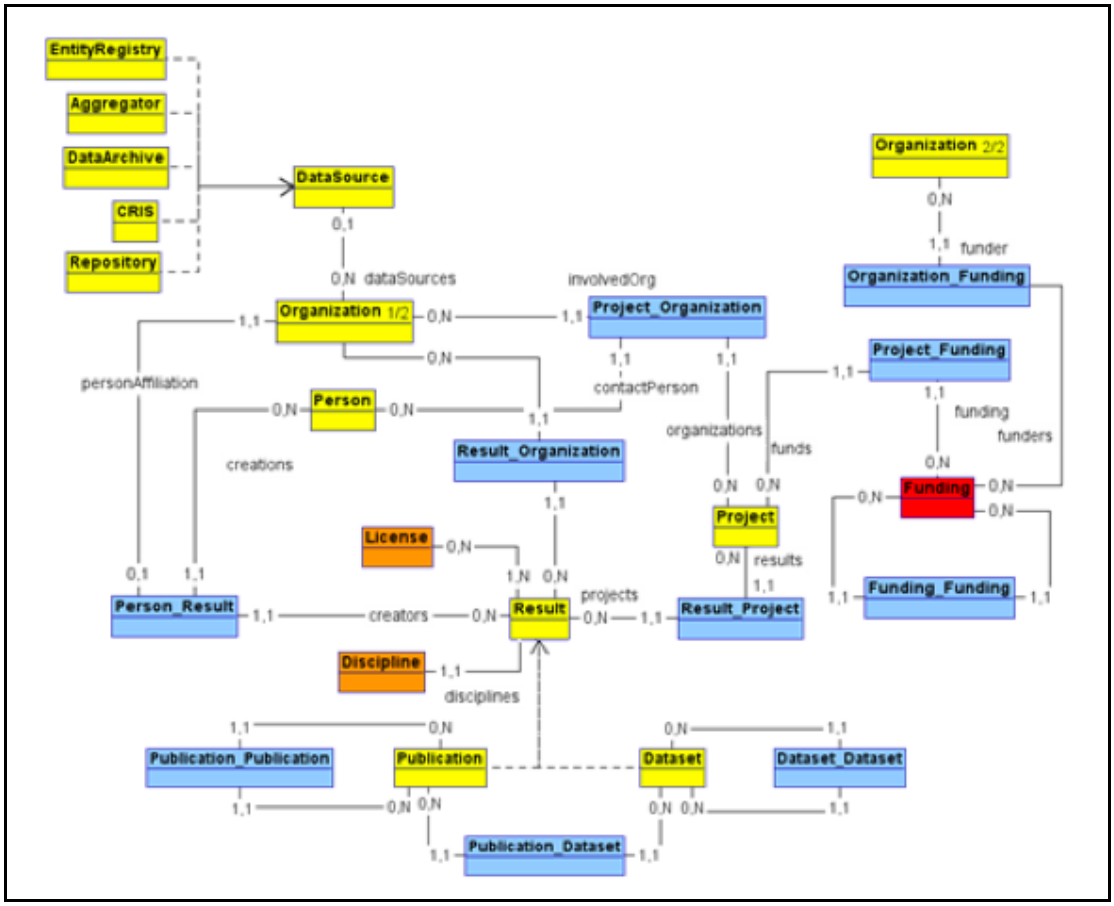


*Figure 2: OpenAIREplus Data Model: Entity-Relationship*

The infrastructure will need to include new mappings for the transformation and cleansing services to cope with the conversion of incoming metadata records (from CRIS systems, dataset archives, and repositories) onto the OpenAIREplus data model.

### Information Space curation and enrichment

Concerning data enrichment, OpenAIREplus research activities invests a lot of effort in automatic inference of relationships between entities in order to compensate for the absence of inter-entity associations in the aggregated data. This absence is mainly due to the multi-disciplinary nature of the incoming data, since information across different communities is rarely connected, and to the general lack of best practices to maintain such associations even in the context of a single discipline; for example, dataset archives and repositories do not generally keep associations from datasets to publications and vice versa (there are exceptions though, e.g., PANGAEA data archive[6], UKPMC literature archive[7]). Relationships not only allow for better search and navigation patterns, but also provide richer context for entities, thereby facilitating the disambiguation process described below, e.g., facilitating the process of determining whether two documents were authored by the same person.

To this aim, the OpenAIREplus infrastructure will feature several knowledge discovery modules for inferring relationships between entities such as documents, datasets, authors, topics, funding sources, etc. Inferred relations will be persisted in the Graph Storage Service (based on HBase distributed storage and Hadoop distributed computing framework) and will be available as five-star[8] Linked Open Data (see Figure 3) via a SPARQL end-point.

Text mining techniques will be employed to spot references to other documents and data sets (represented by *dc:references* predicates,) or to inspect acknowledgments for mentions of funding sources (*oap:funding* predicates.) Coarse-grained document classification, yielding relations between documents and subjects (*dc:subject* predicates) will serve as a prerequisite for fine-grained analysis of the most similar documents (*dc:related* predicates.) Analysis of document texts will be supplemented by association mining based on logs of user activity. Usage pattern analysis will, in particular, discover further valuable relations between documents (*dc:related* predicates) that are not apparent from their texts.

Finally, the infrastructure will offer "end-user behavior analysis" services designed to identify relationships between entities based on user search, navigation, and access patterns.

Concerning data curation, the infrastructure will enhance publication and authors de-duplication services by allowing the re-use of context information, e.g., the publication of an author, to run more accurate record matching algorithms. Furthermore, it will include Web 2.0-like services, which enable registered users to leave data curation "feedback" while searching and browsing the OpenAIRE information space from the portal. Data curators will be able to validate, apply, and undo such feedback from administrative interfaces.

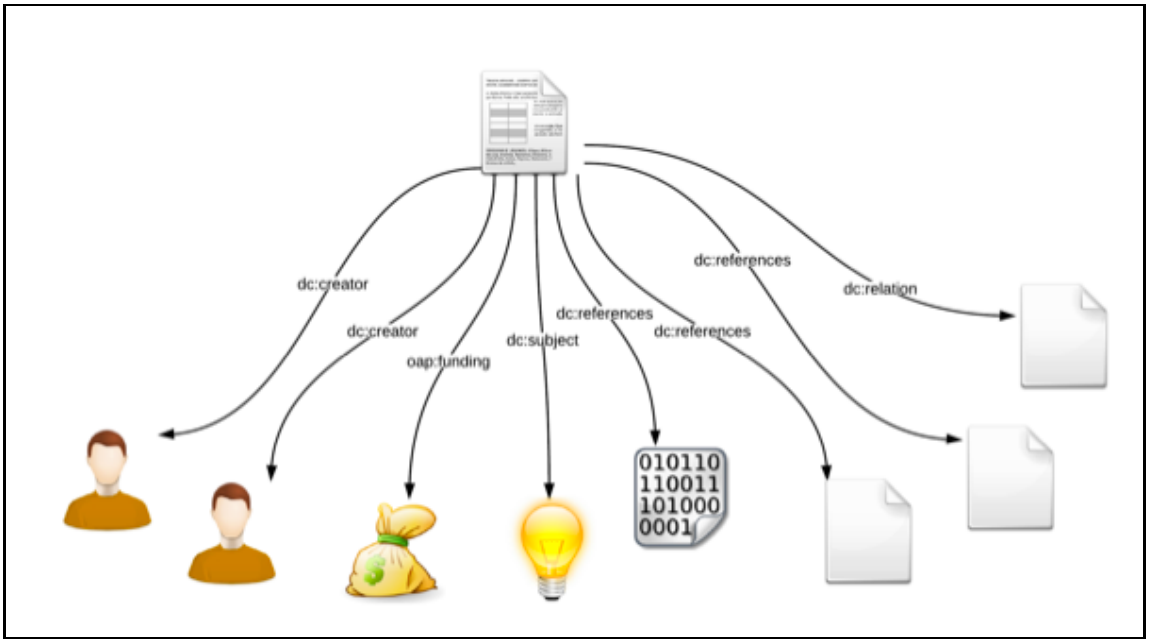


*Figure 3: Relations inferred from a single document, presented as a fragment of a larger semantic network.*

**Access to the Information Space**

The OpenAIREplus portal will provide end user functionalities to manage "enhanced publication" objects (creation, editing, deletion), to "claim" project results (datasets and publications), and to submit feedback on how to improve the information space. Finally, it will export its content through OAI-PMH under the CERIF-EuroCRIS metadata format.

## 4. Conclusions

The OpenAIRE infrastructure was conceived as a "horizontal" EC initiative, i.e., cross-discipline and cross-

country, aimed at connecting research outcomes from scientific "vertical" infrastructures, i.e., discipline specific. The infrastructure will promote the Open Access culture for both publications and datasets; speed-up the research life-cycle by identifying and providing all actors in the scholarly communication chain with new meaningful links between information objects of different types; and leverage multi-disciplinary research. The infrastructure data model (CERIF-based) and architecture (D-NET powered) are designed and implemented to cope with the expected evolution of data representation and functional requirements. In the future, the infrastructure will move towards realizing (or integrating from existing "vertical" infrastructures) discipline-specific services, for example, services focusing on identifying tailored relationships between given publication and dataset types, as well as services capable of managing or accessing such objects, taking advantage of their discipline-specific features.

## 5. Acknowledgements

The work presented in this paper would not be possible without the key contribution of the OpenAIRE technical team members: M. Artini, C. Atzori, A. Bardi, S. La Bruzzo, M. Mikulicic (ISTI-CNR, Italy), L.Nielsen, S. Kaplun (CERN, Switzerland), A. Nowinski, M. Kobos (ICM, Poland), T. Papapetrou, A. Lempesis, K. Iatropoulou, H. Dimitropoulos (University of Athens, Greece), M. Imialek, M. Loesch (Bielefeld University, Germany); and the experience of D. Castelli, W. Horstmann, Y. Ioannidis, and W. Sylwestrzak. Research partially supported by the European Commission as part of the projects OpenAIRE (FP7-INFRASTRUCTURES-2009-1, Grant Agreement no. 246686) and OpenAIREplus (FP7-INFRA-2011-2, Grant Agreement no. 283595).

## 6. References

[1] *European Commission Open Access Pilot*, http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-pilot_en.pdf

[2] *ERC OA mandate*, http://erc.europa.eu/sites/default/files/document/file/erc_scc_guidelines_open_access.pdf

[3] Rettberg, N., & Schmidt, B. (2012) "Repository communities in OpenAIRE: Experiences in building up an Open Access Infrastructure for European research", Open Repositories 2012

[4] Manghi, P., Mikulicic, M.: PACE: A general-purpose tool for authority control. Metadata and Semantic Research pp. 80—92 (2011), http://dx.doi.org/10.1007/978-3-642-24731-6_8

[5] CERIF Data Model, EuroCRIS, http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

[6] PANGAEA data archive of earth & environmental sciences that links to scientific literature, http://www.pangaea.de/

[7] UKPMC archive of life sciences journal literature that links to bio entity databases, http://ukpmc.ac.uk/

[8] *Linked Data design issues*, http://www.w3.org/DesignIssues/LinkedData.html

## About the Authors

**Paolo Manghi** is a Researcher at the Networked Multimedia Information Systems laboratory of the Istituto di Scienza e Tecnologie dell'Informazione (ISTI), Consiglio Nazionale delle Ricerche, Pisa, Italy. His interests are in the fields of Data Models for Digital Libraries, Types for Compound Objects in Digital Repositories, Digital Library Systems and Services, Service-Oriented Infrastructures for Digital Libraries, Database Systems, Type systems for XML languages, Query languages for XML data, and XML P2P database systems. He is currently working as a software architect and researcher for the development of the Digital Library and Data infrastructures for the European

Commission projects OpenAIRE, OpenAIREplus, EFG1914, and HOPE.

**Lukasz Bolikowski** is an Assistant Professor at the Interdisciplinary Centre for Mathematical and Computational Modelling at University of Warsaw (ICM UW). He defended his PhD thesis on semantic network analysis at Systems Research Institute of Polish Academy of Sciences. He is a leader of a research group focusing on scalable knowledge discovery in scholarly publications. He has contributed to a number of European projects, including DRIVER II, EuDML, OpenAIREplus. Earlier at ICM UW he specialized in construction of mathematical models and their optimization on High Performace Computing architectures.

**Natalia Manold** is a Senior Software Engineer holding a B.Sc. in Physics from the University of Athens, Greece, and an M.Sc. in Electrical and Computer Engineering from the University of Wisconsin at Madison, USA. Her professional experience consists of several years of employment as a Software Engineer, Software Architect, Information Technology Administrator, and Information Technology Project Manager by companies in various Information Technology sectors in the US and Greece. The systems she has designed and implemented include specialized ETL tools, biotechnology and genetic applications, embedded financial monitoring systems, and heterogeneous data integration systems. She has also participated and technically managed several R&D projects funded by the European Union (DIAS, DRIVER, DRIVER-II, OpenAIRE, CHESS) or by the national government.

**Jochen Schirrwagen** is a research fellow at Bielefeld University Library, Germany. He has a scientific background in Computer Engineering. He worked for the Digital Peer Publishing Initiative for Open Access eJournals at the academic library center "hbz" in Cologne (2004-2008). Since 2008 he is working for DFG and EU funded projects, like DRIVER-II, OpenAIRE and OpenAIREPlus. Jochen is interested in the application of metadata information using semantic technologies for the aggregation and contextualization of scientific content.

**Tim Smith** leads the CERN group that develops, installs and maintains instances of Invenio, the CERN Open Source Digital Library system. He is heavily involved in initiatives to drive digital archives at the institutional and subject level and to populate them with content of a broad range of media types. Dr. Smith is one of the driving forces behind INSPIRE. His group also assures many other uninterrupted IT services for the 10,000 strong CERN user community. Prior to this task, he led teams responsible for computing farm management and physics data management, innovating in both fields to drive the capabilities up by orders of magnitude. He holds a PhD in Physics and performed research at the CERN LEP accelerators for 10 years before moving to IT. He was a work-package manager of the EU DataGrid project, the forerunner of EGEE.



PRINTER-FRIENDLY FORMAT Return to Article