ARTICLE
EXCERPTED
FROM:

INFORMATION STANDARDS QUARTERLY
FALL 2013 | VOL 25 | ISSUE 3 | ISSN 1041-0031

ISQ

TOPIC

DATA CURATION

DATA CURATION ISSUES IN THE CHEMICAL SCIENCES

DATA CURATION IN THE OPENAIRE SCHOLARLY COMMUNICATION INFRASTRUCTURE

PRESERVING THE GREY LITERATURE EXPLOSION: PDF/A AND THE DIGITAL ARCHIVE

ENSURING THE LONG TERM IMPACT OF EARTH SCIENCE DATA THROUGH DATA CURATION AND PRESERVATION

NISO
How the information world
CONNECTS

Data Curation in the OpenAIRE Scholarly Communication Infrastructure

JOCHEN SCHIRRWAGEN, PAOLO MANGHI, NATALIA MANOLA, LUKASZ BOLIKOWSKI, NAJLA RETTBERG, AND BIRGIT SCHMIDT

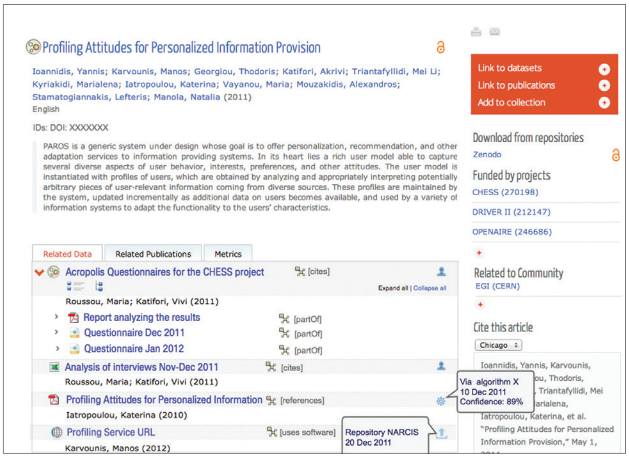
Scholarly communication is currently at a new phase where researchers' published results are more optimally shared, discovered, validated, and reused when they are exposed in their full context. This means that they are best accompanied by all the relative information that provides an insight and capacity to translate the research process and activities that have taken place.

Such information may be program funding (projects), associated datasets, related publications, citations, institutional affiliation, and also a different range of metrics indicating the scientific impact. To best exploit this range of research objects, they can manifest themselves in "enhanced publications." OpenAIRE[1] —the European Union initiative for an Open Access Infrastructure for Research in Europe—supports this enhanced form of open scholarly communication and provides access to the research output of European funded projects and open access content from a network of institutional and disciplinary repositories. Based on a series of European projects (DRIVER and DRIVER-II), Europe has built an extensive network of institutional repositories that follow a common path for interoperability via the publication and consistent use of guidelines. In its current status, OpenAIRE has aggregated more than eight million bibliographic records from over 400 providers of publications which are complemented by research data and research information, where and when the information is available.

OpenAIRE moves beyond the traditional publications aggregator by interconnecting entities related to scholarly communication (currently limited to publications, research data, people, organizations, and their data sources) allowing users to navigate alongside a rich information space graph (Figure 1). It provides individual and aggregated statistics based on different facets, attributes, and linkages of the above entities. As of December 2010, OpenAIRE has been a key service of the EC's Seventh Framework Programme, at present extending to cover other funders and eventually all the European Research Area. As such, OpenAIRE users, i.e., researchers, funding organizations, and third-party services consuming information, expect to find consistent and qualitative metadata. This places data curation as a high priority, employing a series of activities both on the technical

CONTINUED »

[1] The infrastructure is currently funded by the EU-FP7 Grant ID 283595 (OpenAIREplus) at www.openaire.eu



Profiling Attitudes for Personalized Information Provision

Ioannidis, Yannis; Karvounis, Manos; Georgiou, Thodoris; Katifori, Akrivi; Triantafyllidi, Mei Li; Kyriakidi, Marialena; Iatropoulou, Katerina; Vayanou, Maria; Mouzakidis, Alexandros; Stamatogiannakis, Lefteris; Manola, Natalia (2011)
English

IDs: DOI: XXXXXXX

PAROS is a generic system under design whose goal is to offer personalization, recommendation, and other adaptation services to information providing systems. In its heart lies a rich user model able to capture several diverse aspects of user behavior, interests, preferences, and other attitudes. The user model is instantiated with profiles of users, which are obtained by analyzing and appropriately interpreting potentially arbitrary pieces of user-relevant information coming from diverse sources. These profiles are maintained by the system, updated incrementally as additional data on users becomes available, and used by a variety of information systems to adapt the functionality to the users' characteristics.

Related Data | Related Publications | Metrics

Acropolis Questionnaires for the CHESS project [cites]
Expand all | Collapse all
Roussou, Maria; Katifori, Vivi (2011)
› Report analyzing the results [partOf]
› Questionnaire Dec 2011 [partOf]
› Questionnaire Jan 2012 [partOf]

Analysis of interviews Nov-Dec 2011 [cites]
Roussou, Maria; Katifori, Vivi (2011)

Profiling Attitudes for Personalized Information [references]
Iatropoulou, Katerina (2010)

Profiling Service URL [uses software]
Karvounis, Manos (2012)

Via algorithm X
10 Dec 2011
Confidence: 89%

Repository NARCIS
20 Dec 2011

Link to datasets ⊕
Link to publications ⊕
Add to collection ⊕

Download from repositories
Zenodo

Funded by projects
CHESS (270198)
DRIVER II (212147)
OPENAIRE (246686)
+

Related to Community
EGI (CERN)
+

Cite this article
Chicago ⇕

Ioannidis, Yannis, Karvounis,
ou, Thodoris,
Triantafyllidi, Mei
arialena,
Iatropoulou, Katerina, et al.
"Profiling Attitudes for Personalized Information Provision," May 1,

Figure 1: A publication view in OpenAire

CONTINUED »

and networking levels. Data curation in OpenAIRE differs from similar activities in a data archive as it is mainly applied on the metadata level, needing to comprise heterogeneous data sources and is backed by complex organizational and technical support structures. It can also comprise of a range of actions and workflows, as will be demonstrated below.

OpenAIRE, as the *pan*-European infrastructure of *all* scientific results, faces two major challenges: diversity of scientific data and diversity of cultural approaches for good practices that span many national borders. To understand how expert research communities facilitate data curation mechanisms and carry out linking to supplemental data within their distinct fields, as well as the challenges that might be posed in creating a cross-disciplinary infrastructure, OpenAIRE has worked closely with three scientific partners: DANS for social sciences, EBI-EMBL for life sciences, and BADC for earth observation data. It has explored possible commonalities in practices and has looked at how to incorporate discipline-specific vocabularies and indexes that may be used in a horizontal data-sharing environment. To understand national and local needs, OpenAIRE's infrastructure is complemented by an active networking effort which supports users and consumers of the system: the mainly research library-led community comprises a network of 33 pan-European advocacy offices which promote and consolidate open access policies and benefits. This is via a range of dissemination activities and exchange of information at both national and European levels to all stakeholders, such as research libraries, publishers, and repository communities. Supporting researchers in publishing in open access is also a key task of the national offices, and this has now extended to the concept of open data, and an awareness of sharing and making as accessible as possible all research outcomes, such as the data collected during the research process.

This article outlines the curation activities conducted in the OpenAIRE infrastructure, which employs a multi-level, multi-targeted approach: the publication and implementation of interoperability guidelines to assist in the local data curation processes, the data curation due to the integration of heterogeneous sources supporting different types of data, the inference of links to accomplish the publication research contextualization and data enrichment, and the end-user metadata curation that allows users to edit the attributes and provide links among the entities.

Guidelines are the glue for interoperability

Raising visibility of global research output is possible via interoperability between data infrastructures. Early on, it became evident that repositories need a common ground to work on their interoperability issues, both on the syntactical and semantic layer. Publication (literature) repositories have had a head start with OAI-PMH, while the DRIVER guidelines have become an international standard frequently enforced by many policy makers. Guidelines address different aspects of the publishing and exporting process and are very much related to proper data curation techniques at the archiving level. They recommend the use of controlled vocabularies and formats where necessary (e.g., publication/data type, languages, dates, access mode, and rights), the proper encoding of information about funding bodies and project information, the use of persistent identifiers (e.g., DOIs, author IDs), the optimal use of transfer protocols, etc. They extend to the documentation of good practices for properly housed datasets with accepted Research Data Management plans in place, as well as the effective advocacy about the benefits of exposing metadata for effective harvesting.

OpenAIRE's guidelines are a central tool dedicated to content providers for becoming compatible with the OpenAIRE infrastructure.

Working with related initiatives (COAR, DataCite, EuroCRIS), OpenAIRE has produced three sets of guidelines classified under the following schemes:

1. ***Literature Repository*** **Guidelines exposing OAI-DC**
   This assists repository managers to improve the discoverability of research output and meet EC Open Access requirements for publications to authors, and eventually the requirements of other (national or international) funders with whom OpenAIRE cooperates.

2 *Research Data Repository* **Guidelines exposing metadata, extending the DataCite Metadata Schema**
This assists data archive managers to standardize the commonly used metadata attributes. These are mainly related to the provenance and links to other entities (literature, funding), used as a stepping-stone for a linked data infrastructure for research and the facilitation of enhanced publications.

3 *CRIS applications* **supporting a CERIF compliant XML-profile (to be completed in fall 2013)**
This essentially offers a mapping of research information systems to repository metadata, supplementing links among the scholarly communication entities.

As research data now becomes a prominent player in scholarly communication activities, and given the incredibly varied nature of data repositories (from specialized ones serving discipline specific communities to general purpose ones established at research institutions), OpenAIRE is intensifying its efforts on the guidelines front and aims to produce a toolkit of best practices, including a list of pointers for a complete reference for those planning to set up a data repository.

Supportive services for data curation

The technical infrastructure of OpenAIRE is based on the open and scalable software D-NET that offers a rich set of data management services and supports layered functionalities for the aggregation, advanced management of scientific information, and interaction with end users and third-party service providers. It relates objects from a diverse range of data sources using a variety of ingesting workflows. The data flow (Figure 2) is divided into interdependent phases ranging from automated processes to import data from the affiliated content providers (harvesting for repositories, APIs for registries), data normalization, and de-duplication, to semi-automated data enrichment processes for knowledge inference (linking and classification), to manual interventions by end users to edit metadata and provide context associated with the research output (claims and feedbacks). These phases are well orchestrated, each associated with different data curation mechanisms as explained in the next sections. Furthermore, the data is enriched with provenance information that allows distinguishing in which phase the data was inserted and by which agent (human or service). Such tracking allows the phases to be rolled back and repeated without losing the results obtained in the previous phases.

CONTINUED »

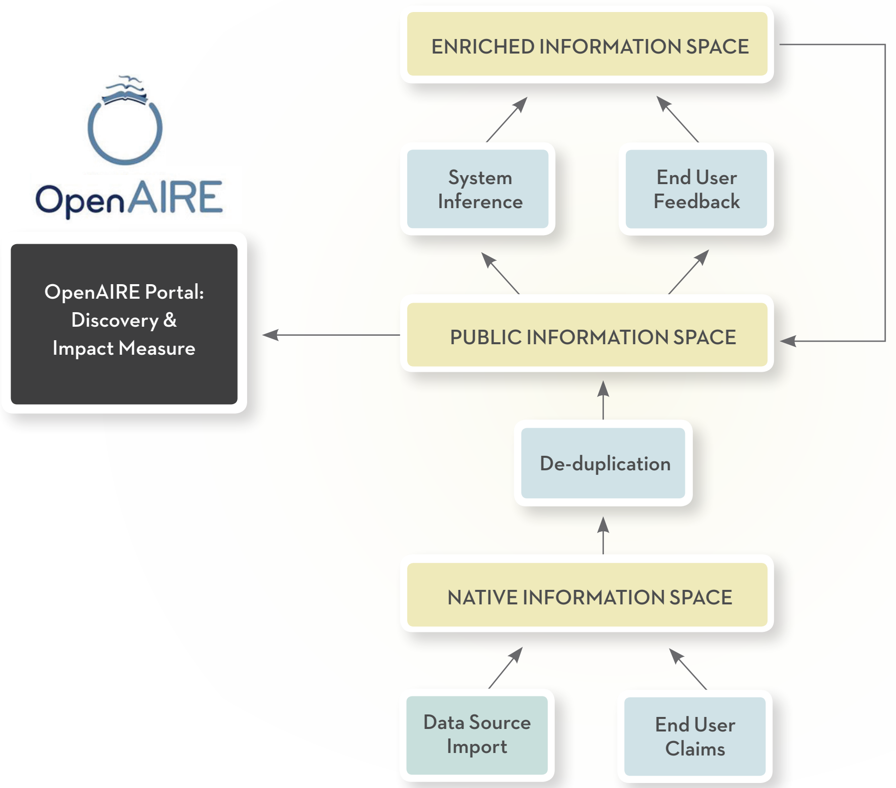


Figure 2: Data flow and population of OpenAIRE's information space

Data sources and data collection

Content is ingested from literature and research data repositories, CRIS systems, and funder databases. It synchronizes with authoritative repository and author registries and resolver systems. The purpose is to "contextualize" publications and datasets with curated information. Often, these data sources provide interlinked content, which has already been curated and verified as authoritative and trusted. The use of authority files plays an important role in all ingestion workflows, covering the list of registered repositories and policies,[2] projects funded by research organizations,[3] organization names and their typology, and author IDs. The lack of authority files poses a great challenge, so OpenAIRE partners with initiatives that maintain many of the authority files used internally, once again illustrating the need for an eco-system of interoperable and interdependent infrastructures and data services.

The first phase of data curation for repositories takes place during their registration in the infrastructure through a validation process. An independent service, the validator, is deployed and used by repository managers to validate both the optimal practices of the OAI-PMH transfer protocol and the correct use of metadata fields, providing a first estimation of the overall metadata quality and compatibility level to the OpenAIRE guidelines. Through an iterative, most often human-mediated process, repository managers are advised how to best comply with the guidelines. Each repository is assigned a score, which is presently used only for internal purposes, but is envisioned to be used in a rewarding mechanism within the infrastructure. The validator service has proven quite valuable over its five years of operation from the DRIVER network era, has been used by over 450 repositories, and is the designated tool in some national repository infrastructures (e.g., Spain, Poland, Slovenia, and Argentina). It is a flexible, autonomous, and configurable service based on extensible rules and can validate content against different sets of guidelines. It is now being extended to accommodate different types, more complex data catalogs,[4] and transfer protocols.

Once a repository is registered and validated, there is a continuous process of administration and maintenance of registered data sources. This is due to a number of factors; for example, new policies may be applied within the repository, including changes in metadata practices; repository software may be updated to a new release; or the location of repositories may change, disappear, or become merged with other repositories. This requires constant validation, which is also embedded in the OpenAIRE workflows, accompanied by human intervention and communication with the repository managers.

Data normalization and entity de-duplication

Once the metadata from repositories (publications and research data) is ingested into OpenAIRE, a normalization process takes place consisting of the steps outlined below.

First, the metadata gets normalized and is transformed into a harmonized internal representation. The transformation is based on XSLT (Extensible Stylesheet Language), which can be configured and adjusted individually to each repository, since practice has shown that repositories consistently use the same formats and values. During the normalization procedure, the set of controlled vocabularies, published in the OpenAIRE guidelines, is utilized to effectively map and relate all corresponding publication or research data attributes. Additional processes are enabled to fill in missing values, a case often encountered in data curation processes. These currently include language detection mechanisms, retrieval of alternative identifiers through the use of external APIs (e.g., PubMed/PMC IDs or DOIs), and retrieval of links to related data as published from discipline-specific community services. Furthermore, each metadata record is supplemented with its provenance data and is assigned an internal identifier.

Second, a de-duplication process takes place, which is quite an important step for the infrastructure outcome since OpenAIRE produces statistics that are frequently used in assessing impact of funding and are even used by high-level EC ranks while drafting future policies. De-duplication is not only important because of the variance of integrity of the metadata in the records and variant forms of spelling of attribute values (e.g., titles, author names, funding projects, subjects, institution names), but also because collection of data from various sources implies duplicates of records referencing the same publication, as these may be deposited in different locations by the same or different persons or processes. OpenAIRE attempts to integrate duplicate records by the de-duplication of person names and publication records (Figure 3).

[2] List of institutional and thematic repositories obtained from OpenDOAR and the list of research data repositories is obtained from Registry of Research Data Repositories (re3data).

[3] OpenAIRE currently collects from the EC and Wellcome Trust databases (via PMC) but is extending to other European funders.

[4] It is also configured for DataCite XML Schema, and is currently being updated for use in the ESPAS data infrastructure (Near Earth Space Data) with the OGC catalogue service protocol (Open Geospatial Consortium).

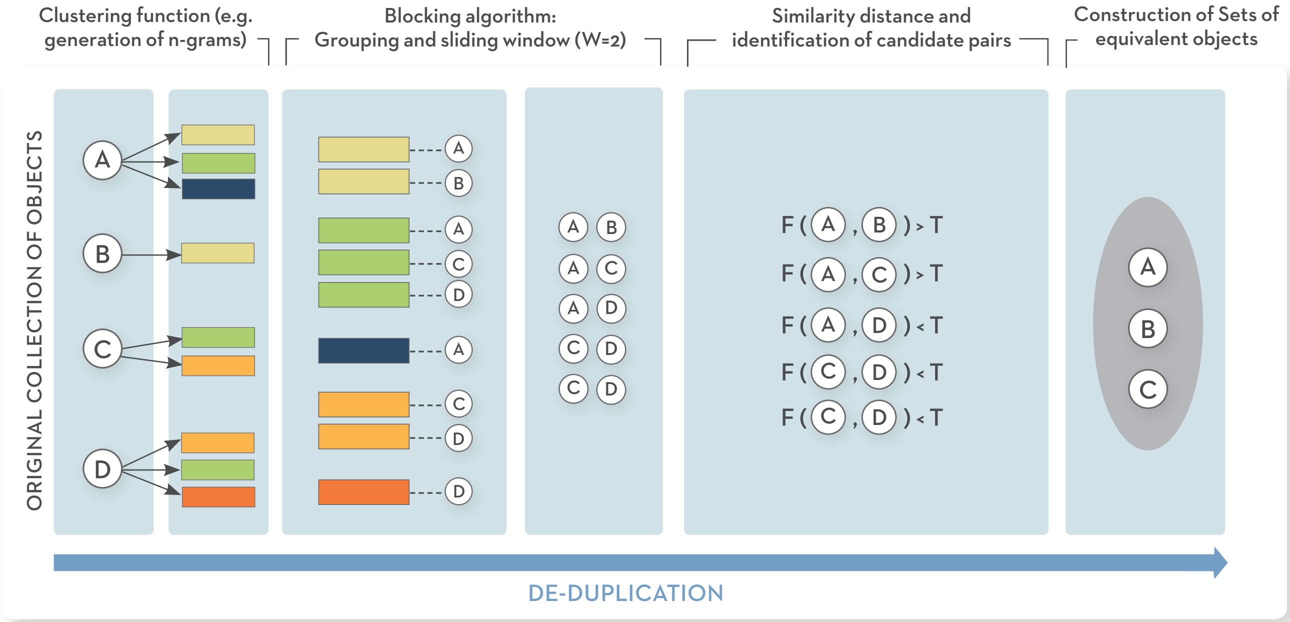

Clustering function (e.g. generation of n-grams)
Blocking algorithm: Grouping and sliding window (W=2)
Similarity distance and identification of candidate pairs
Construction of Sets of equivalent objects
ORIGINAL COLLECTION OF OBJECTS
A
B
C
D
F(A,B) > T
F(A,C) > T
F(A,D) < T
F(C,D) < T
F(C,D) < T
DE-DUPLICATION

Figure 3. Logical steps of the de-duplication process

Two types of de-duplication actions are performed using similarity matching:

1. **Entity disambiguation,** when distinct entities with different identities are represented by distinct objects with different identifiers; and
2. **Entity association,** by interlinking different objects to indicate they are different versions or variations of the same entity.

Knowledge Inference: semi-automated contextualization

The data in OpenAIRE, as collected from data providers and normalized in the initial parts of the workflow, contain tremendous amounts of knowledge "hidden" both in the metadata and the full text files of publications. The process of data collection is complemented by web-crawling of metadata on research publications and full-text downloads[5] feeding an OpenAIRE subsystem, the Information Inference Service (IIS), that analyzes available data and enriches its information space with knowledge inferred from the contents. This includes inference/mining algorithms to analyze the OpenAIRE graph of objects or the documents (e.g., PDFs, XMLs) associated to them, in order to identify relevant relationships between such objects, new objects, or object property values. Knowledge inference is fully automated and the IIS is a framework, scalable to accommodate millions of publications and flexible to extend to different or additional inference modules. Its components make use of text mining and statistical machine learning approaches and consist of a set of independently developed modules/services, each performing a specific extraction or inference task.

The main services currently under operation or development are:

1. **Identification of projects:** text mining algorithms detecting funding schemes from an arbitrary number of funding bodies and institutions using pattern matching and contextual information to provide confidence level and filter out the false matches
2. **Identification of data citations:** by text mining of known identifier schemes (e.g., DOIs, PDB IDs, Genbank IDs, PubMed/PMC IDs, etc.) combined with metadata similarity matching
3. **Identification of persons:** featuring extraction algorithms to identify the names of authors together with their affiliation
4. **Subject classification:** supervised classification to classify text into a set of pre-defined class/publication labels using taxonomies (e.g., arXiv subject classification, Dewey Decimal Classification)
5. **Clustering:** probabilistic topic modeling aiming to discover and annotate with thematic information and identify useful patterns, similarities, and communities in related multi-dimensional linked data and attributes

CONTINUED »

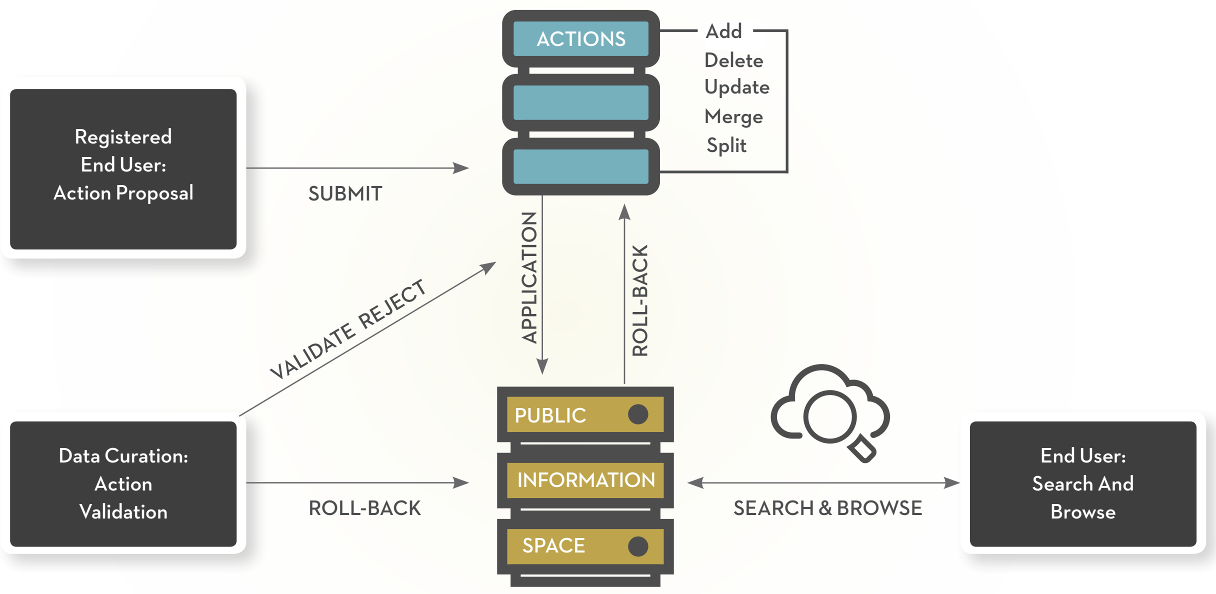


Figure 4: End user feedback workflow

Automatically inferred information may be invalidated by human data curators while inference is performed periodically (weekly) in a batch mode. In addition, OpenAIRE keeps track of data provenance (in this case which algorithm the enrichment came from), and is able to assess the "level of trust" of every single piece of information.

End user feedbacks

Data curation in OpenAIRE is complemented by communication with end users. The current operation supports "claims"[6] which are statements from users who can, through the portal, select results (internal or external) and link to EC or other funder projects. Upcoming releases[7] will provide registered users with enhanced services for submitting "advice" on how to improve various facets of the information space as this is formed after the ingestion and curation by the preceding processes. Specifically, users will be able to provide suggestions on how to correct metadata fields and enrich objects through the assignment of projects, related publications, citations of datasets, etc. They will be able to group scientific results into *aggregations* or *collections* for providing additional context related to their research activities. All these end user "actions" are preserved and, before being applied, need to be validated by OpenAIRE data curators.

Actions may involve:

» Updates of properties to individual objects in the information space (metadata edits)

» Addition of objects and addition/removal of relationships between such objects (claiming)

» Merging of objects considered to be descriptions of the same real-world object

» Splitting of one object into two or more objects

As illustrated in Figure 4, actions are submitted by users in a "pending" status, i.e., not yet visible as part of the public information space, accompanied by a corresponding trust level. Data curators, i.e., designated users with advanced privileges, will be notified of new pending actions and will be able to "validate" actions and pass them into the information space in a visible state. Data curators will also be able to "roll-back" such actions, that is to remove them and all their subsequent consequences on the information space. Given their "expert" role, data curators can directly submit "validated actions."

[5] Special agreements are made with the data providers to eliminate IPR issues.
[6] ~ 2000 claims related to EC publications have been recorded in the OpenAIRE portal from December 2011.
[7] The implementation of these services is a work in progress and will be public in early 2014.

Conclusion

Aggregation and enrichment of "data" in a participatory knowledge infrastructure such as OpenAIRE involves specific data curation activities additional to those at the source level (repository, data archive). Collecting data of diverse types from heterogeneous sources relies on a level of implicit "trust" (trusted actions) in the harvested metadata regarding accuracy, persistence, and accessibility. The agreement on and implementation of guidelines for the use of metadata schemes and transfer protocols are essential to ensure a uniform interpretation of information and to foster interoperability across infrastructures. OpenAIRE represents an important first step towards a curated information space for scholarly communication, putting foundations through its widely used services. Nevertheless, the heterogeneity and the unprecedented size of the data involved, as well as the diversity of policies and approaches in the European repository landscape raise considerable challenges that need to be addressed in an ongoing effort.

| IP | doi: 10.3789/isqv25no3.2013.03

JOCHEN SCHIRRWAGEN (jochen.schirrwagen@uni-bielefeld.de) is responsible for data aggregation in OpenAIRE and is task coordinator on Subject Specific Pilots for Enhanced Publications, Bielefeld University Library. **PAOLO MANGHI** (paolo.manghi@isti.cnr.it) is OpenAIRE Technical Manager and is affiliated with the Institute of Information Science and Technologies of the Italian National Research Council. **NATALIA MANOLA** (natalia@di.uoa.gr) is OpenAIRE Project Manager and is affiliated with the University of Athens and ATHENA Research Center. **LUKASZ BOLIKOWSKI** (l.bolikowski@icm.edu.pl) leads a research group on scalable knowledge discovery in scholarly publications and is affiliated with the University of Warsaw Interdisciplinary Centre for Mathematical and Computational Modelling. **NAJLA RETTBERG** (najla.rettberg@sub.uni-goettingen.de) is OpenAIREplus Scientific Manager, University of Goettingen. **BIRGIT SCHMIDT** (bschmidt@sub.uni-goettingen.de) is OpenAire Scientific Manager, University of Goettingen.

RELEVANT LINKS

British Atmospheric Data Centre
badc.nerc.ac.uk/home/index.html

Confederation of Open Access Repositories
www.coar-repositories.org

Data Archiving and Networked Services (DANS)
www.dans.knaw.nl/en

DataCite
www.datacite.org

DRIVER (Digital Repository Infrastructure Vision for European Research)
www.driver-repository.eu/

DRIVER guidelines
www.driver-support.eu/managers.html#guidelines

D-NET
www.d-net.research-infrastructures.eu/

DataCite Metadata Schema Repository
schema.datacite.org/

EuroCRIS
www.eurocris.org

European Bioinformatics Institute (EMBL-EBI)
www.ebi.ac.uk/

Near Earth Space Data
www.espas-fp7.eu

OpenAIRE
www.openaire.eu

OpenAire Literature Repository Guidelines
guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_Literature_repositories

OpenAire Research Data Repository Guidelines
guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_Data_Archives

OpenAire Validator for repositories
validator.openaire.eu

OpenDOAR
www.opendoar.org

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
www.openarchives.org/pmh/

Open Geospatial Consortium, Catalogue Service Protocol
www.opengeospatial.org/standards/specifications/catalog

Registry of Research Data Repositories
www.re3data.org

Seventh Framework Programme for Research (FP7)
cordis.europa.eu/fp7/home_en.html

XSLT (Extensible Stylesheet Language)
www.w3.org/Style/XSL/