OPEN SCIENCE AS A SERVICE WORKSHOP: TOOLS FOR RESEARCH COMMUNITIES

Paolo Manghi and Alessia Bardi

The aim of the Open Science as a Service workshop was to create a forum where research communities especially interested/relevant to open science publishing paradigms can compare, confront and align their practices by identifying common patterns and methodologies. The result have become important feedback and validation to the architecture, information models, and vision of the services to be provided by OpenAIRE in support of Open Science. The workshop counted around 60 participants, was very interactive and structured as follows:

Welcome and Introduction

- "Open Science Publishing Challenges - an overview", Paolo Manghi, OpenAIRE, ISTI-CNR (IT)\

Community perspectives of Open Science

- "A Structural Biology view of Open Science", Chris Morris, STFC (UK)
- "An Ecology and Evolution view of Open Science", Antica Culina, KNAW (NL)
- "The ELIXIR view of Open Science", Thanasis Vergoulis, Athena Research and Innovation Centre (GR)

OpenAIRE-Connect Services for Open Science as-a-Service

- "The Research Community Dashboard", Paolo Manghi, OpenAIRE, ISTI-CNR (IT)
- "The Catch-All Broker Service", Alessia Bardi - ISTI-CNR (IT)

Open Science Publishing Challenges - an overview

Everyone is talking about OpenScience. Scientists and organizations see it as a way to speed up, improve quality, and more effectively reward research activities, while funders and ministries see it as a means to optimize cost of science and leverage innovation. Open Science is an emerging vision, a way of thinking, whose challenges always gaze beyond its actual achievements. Today, the effective implementation of Open Science calls for a scientific communication ecosystem capable of enabling OpenScience publishing principles. The ecosystem should allow research communities to share (for "discovery" and "transparent evaluation") and re-use (for "reproducibility") their scientific results by publishing all intermediary and final research products, beyond scientific literature. Products can be research data, methods, software, workflows, protocols, scripts, algorithms, etc.. They should be deposited in data sources for scientific communication, e.g. institutional repositories, data archives, software repositories, CRIS systems. They should be published together with the semantic links between them. To complete the picture, such ecosystem should support publishing of packages of products (e.g. research objects, enhanced publications, RMap) to allow discovery, evaluation, and reproducibility of combinations of artefacts (e.g. workflows or experiments with input datasets).

Today's scientific communication landscape is far from supporting this vision, mainly due to its inability to (i) support publishing of all kinds of research products: for example, research software publishing workflows are generally not best practice, i.e. no repositories, no dedicated PIDs, no scientific reward; (ii) keep a complete and up-to-date record of research products relationships: for example, publication, data, software repositories and publishers do not keep bilateral links between each other's products, and the links they keep are not in-sync with the evolution of science (e.g. links to new versions of the data, obsolete links); and (iii) find agreements on how to share and publish packages of products: solutions exists but are specific to rather small communities of scientists and, as research software, are not regarded as first-class citizens in the scientific communication domain. De facto, today's scientific communication ecosystem lacks tools and practices for engaging research communities at adopting the aforementioned novel Open Science publishing principles, even when researchers are already in the position of publishing interlinked artefacts and/or product packages.

Community perspectives of Open Science

The three communities presented a variegated panorama, with different degrees of technical maturity, community vision and cohesion, and understanding of Open Science publishing challenges and desiderata.

Structural biology aims to understand the molecular basis of life and for that scientists can rely on well established research infrastructure facilities, services, and best practices. Examples are synchrotrons, electronic microscopes, nuclear magnetic resonance instruments and several other services. The scientific process consists in processing data to identify protein structures by means of software tools and concludes with publishing of articles, datasets (Protein Data Bank), and software used to perform the experiment. Still the community lacks of practices and tools to archive experimental data to provide access to reviewers, miss part of provenance information required to reproduce the experiments, and would benefit from automated acquisition of metadata.

Ecology and evolution focuses on strategic ecological research on individual organisms, populations, ecological communities, and ecosystems. The scientific process in theory is rather simple as it consists in designing the experiment by drafting an hypothesis, data acquisition (e.g. observations, field study), data processing, and finally publishing an article, possibly with “additional material”, inclusive of datasets and software (19% of articles link to source code). In practice, this life-cycle turns out to be very complex due to the heterogeneity of data required for an experiment, which originates in different disciplines, collected and annotated with different methods and are distinct in terms of types, formats, scale, and integration. The community strives for standards and best practices on how to share and access research data and software, and struggles to have Open Science principles endorsed by the community, adopting existing tools in support of it, or considering changes to scientific reward practices.

ELIXIR is an intergovernmental organisation aiming to coordinate, integrate and sustain bioinformatic resources across Europe since 2014. Funded by the EC, includes 22 countries and embraces the effort of 180+ organizations. Scientists in bioinformatics start from an observation draw a theory or hypothesis and verify it by performing in-silico analysis on new data or new analysis on old data. As a result of this process scientists publish articles (often Open Access), algorithms (typically in Python or R), workflows, tools and datasets. ELIXIR provides a repository for algorithms and tools (BioTools) and researchers have good practices in publishing data. Still, the quest to Open Science is ongoing: on the side of publishing, managing multiple versions of datasets is a rather important aspect, not yet fully tackled by best practices and tools; on the side of scientific reward, proper techniques and indicators must be devised, as quantitative measures (e.g. citation counts) are not enough to select the best results out of the scientific paper deluge and to convince scientists at opening up their datasets and tools.

OpenAIRE-Connect Services for Open Science as-a-Service

The OpenAIRE infrastructure's mission is to foster Open Science by advocating its principles and benefits across European countries and research communities and by offering technical services in support of Open Access monitoring, research impact monitoring, and Open Science publishing. Its role in the European Open Science Cloud (EOSC) setting is to provide Research Infrastructures (RIs) with the services required to bridge the research life-cycle they support, where scientists produce research products, with the scholarly communication infrastructure, where scientists publish research products, in such a way science is reusable, reproducible, and transparently assessable. OpenAIRE is fostering the establishment of reliable, trusted, and long lasting RIs, not attending to replace or compete with RI existing services. In this sense, OpenAIRE Open Science services should be intended to compensate the lack of solutions, where these are not available to scientists in their RIs, and a boost to build or upgrade existing solutions to meet Open Science publishing needs.

The first service is the **Research Community Dashboard**. Thanks to its functionality, scientists of RIs can (i) find tools for publishing all their research products, such as literature, datasets, software, research packages, etc. (provide metadata, get DOIs, and ensure preservation of files), (ii) interlink such products manually or by exploiting advanced mining techniques, and (iii) integrate their services to automatically publish metadata and/or payload of objects into OpenAIRE. As a consequence, scientists populate and access an information space of interlinked objects dedicated to their RI, through which they can share any kind of products in their community, maximise re-use and reproducibility of science, and outreach the scholarly communication at large.

The second service is the **Catch-All Broker Service**. Thanks to its functionality, data sources such as institutional repositories, data repositories, software repositories can be notified of metadata records relative to products (datasets, articles, software, research packages) that are "of interest to them", i.e. metadata records that should be in the data source, or "linked to them", i.e. a scholarly link exists between one of the data source product and the identified product. Notifications are sent only to subscribed data sources, following a subscription and notification pattern, and can be delivered by mail, OAI-PMH end-user interfaces, or, currently under investigation, via push APIs (e.g. SWORD protocol), FTP and ResourceSync. The idea behind the service is to disseminate and advocate the principle that scholarly communication data sources are not a passive component of the scholarly communication ecosystem, but rather active and interactive part of it. They should not consider themselves as silos of static products, but rather as hubs of products interlinked with the evolving research ecosystem.

Results and feedback

As a result of the discussions occurred during and at the end of the meeting, it became clear that research communities find it hard to understand and tackle the complexity of Open Science publishing. In the context of RIs part of these challenges is overcome by the existence of a board or legal entity which can issue top-down decisions, or at least recommend them. This strength is particularly evident in the decisions taken about standards for interoperability or dataset publishing policies. Still, reproducibility of science is typically out of such monitors. Where governance does not exist (i.e. the community is not supported by an RI), the issues are even harder to tackle, since communities identify themselves only by the kind of research they perform but find it impossible to converge on such open and complex challenges.

These challenges could be explored in the context of EOSC. For example by identifying models that allow “RI-less communities” to exploit existing RIs and “virtually” build their own RIs out of the services of existing ones (mission of EOSC). Such “virtual” RIs would build and rely on their own “governance”, supporting their own research mission, vision and best practices, but rely on services provided by third parties.

Finally, representative of research infrastructures in the attendance urged OpenAIRE to establish collaborations with RIs to enable automated publishing (push of metadata and links in OpenAIRE) or harvesting from RI sources metadata for datasets, tools, and links to publications. The RIs of ELIXIR, Instruct-WestLife, and OpenRiskNet showed interest in establishing such collaborations. More generally, it became clear how RIs, and more in general communities, need recommendations and counseling to kick-start an Open Science virtuous circle. OpenAIRE could be the “public” entity that will establish “offices” (e.g. Research Community Contact Points) to investigate RI needs and match them with OpenAIRE OSaaS tools to reach integration and mutual benefits.