Horizon 2020 – Outline of a Pilot for Open Research Data

Joint statement by OpenAIRE, LIBER and COAR – 3 July 2013

The European Commission is developing an Open Data Pilot. It will look at research data generated in projects funded under the Horizon 2020 framework, with the aim of stimulating the data-sharing culture among researchers and facilitating both the re-use of information and data-driven science. As organisations with a strong interest in Open Data, *OpenAIRE*, *LIBER* and *COAR* would like to give their views on the current situation and make recommendations for an effective Open Data Pilot.

Overview

Projects funded by the European Commission (EC) and the European Research Council (ERC) produce a wide variety of research data (eg. from experiments, simulations and surveys). These data are typically small sets, scattered across repositories and hard drives throughout Europe. The success of the EC's Open Data Pilot is therefore dependent on support and infrastructures that acknowledge disciplinary approaches on institutional, national, and European levels.

The importance of this underlying support is further clarified when we consider that the management of research data requires more than simply placing data in a repository. To make re-use possible, a sequence of steps (eg. planning, creation, documentation) need to be followed. Since the nature of data varies between disciplines, a selection process is also critical to identify the sets which should be included in the pilot on a per discipline basis.

Research libraries are already mobilising to answer these needs but resource issues such as skills and funding must first be addressed. Similarly, there are several diverse data centres in Europe and world-wide (thematic, national, etc.) that manage research data but better interaction between these centres and across the broader stakeholder community needs to be encouraged.

Cultural attitudes to data management and sharing are another factor to consider. The European research community encompasses a broad range of data management practices and attitudes towards research data sharing, depending on aspects such as the culture within various disciplines and the sensitivity of data.

For all of these reasons, the data-sharing ecosystem must be stimulated and nurtured. The pilot is an excellent opportunity to do this, and has the potential to connect researchers interested in sharing and re-using data with the relevant services within their institutions (library, IT services), data centers and data scientists. The pilot should serve to promote the value of data sharing to both researchers and funders, as well as to forge connections between the various players in the ecosystem.

Within the European Commission's Seventh Framework Program, *OpenAIRE* and *OpenAIREplus* explored a pilot pan-European Open Access infrastructure for publications, resulting in requirements for Horizon 2020 projects. *LIBER* has a long history of supporting libraries in the development of institutional research data management policies and services. Given this

experience, *LIBER*, *COAR* and *OpenAIRE* are in a strong position to provide coordination and support measures to an Open Data Pilot, and to activate research institutions, libraries and other stakeholders in the collection, storage and dissemination of Open Data.

Objectives of the Pilot

The main goal of the pilot is to *facilitate research data registration, discovery, access and re-use*, in particular in the context of Horizon 2020 funded projects across European countries while linking these to international efforts.

To achieve this goal the pilot should:

- Raise awareness among researchers about the need for and the benefits of managing and sharing research data through a pan-European campaign for advocacy of the benefits of data sharing to researchers and for research.

- Encourage libraries, data centres, publishers and other relevant stakeholders to start working with researchers throughout the data life cycle (planning, data creation, documentation, storage, rediscovery, and re-use).

- Identify what is required amongst all stakeholders (researchers, libraries, IT departments) in terms of expertise, knowledge and skills and identify how these gaps will be met.

- Address the creation and interoperability of reliable e-infrastructures allowing the registration, description through the assignment of metadata, storage and curation of research data while linking them to funding information and related research outputs (publications, reports, software etc.).

- Mobilise research libraries to support scientific communities with research data management planning and to have a deeper understanding of their role in dealing with data. Promote their collaboration with affiliated data centers where necessary (national or disciplinary specific ones, like ESFRI nodes).

- Provide the Commission with the necessary evidence to make a rational decision about the next steps necessary to foster research data sharing.

Enabling projects to register, discover, access and re-use research data

Management of research data is a process which has to be addressed throughout the life cycle of data. One core result from the Open Data Pilot should be the support of researchers during this process. This would include the provision of guidance and infrastructure for registration, discovery, access and re-use of data.

#1 What

The Open Data Pilot covers all research data and associated metadata resulting from EC-funded projects, if they serve as evidence for publicly available project reports and deliverables and/or peer reviewed publications. To support discovery and monitoring of research outputs, metadata have to be made available for all datasets, regardless of whether the dataset itself will be available in Open Access. Data repositories might consider supporting the storage of related project deliverables and reports, in addition to research data.

In particular, the pilot covers all projects funded by the European Research Council, and selected Horizon 2020 research areas such as Future and Emerging Technologies, Climate, Biodiversity, Society etc.

#2 Where

All research data has to be registered and deposited into at least one open data repository. This repository should:

- Provide public access to the research data, where necessary after user registration;
- Enable data citation through persistent identifiers (eg. DOI, ARK, PURL);
- Link research data to related publications (eg. journals, data journals, reports, working papers);
- Support acknowledgement of research funding within metadata elements;
- Offer the possibility to link to software archives;
- Provide its metadata in a technically and legally open format for European and global re-use by data catalogues and third-party service providers based on wide-spread metadata standards and interoperability guidelines.

Data should be deposited in trusted data repositories, if available. These repositories should provide reliable long-term access to managed digital resources and be endorsed by the respective disciplinary community and/or the journal(s) in which related results will be published (e.g., Data Seal of Approval, ISO Trusted Digital Repository Checklist).

Registries of data repositories are re3data (www.re3.data.org) and Databib (www.databib.org).

#3 When

Research data related to research publications should be made available to the reviewers in the peer review process. In parallel to the release of the publication, the underlying research data should be made accessible through an Open Data repository (see #2).

If the project has produced further research datasets (i.e. not necessarily related to publications) these should be registered and deposited as soon as possible, and made openly accessible as soon as possible, at least at the point in time when used as evidence in the context of publications.

#4 How

The use of appropriate licenses for Open Data is highly recommended (e.g. Creative Commons CC0, Open Data Commons Open Database License).

#5 Who

Responsibility for the deposit of research data resulting from the project lies with the project coordinator (delegated to project partners where appropriate).

#6 Supporting actions

A number of supporting actions will ensure the Open Data Pilot achieves its aims.

A pan-European advocacy campaign to researchers, funders, universities, librarians and IT support staff is needed to sell the benefits of data sharing and re-use to the researcher, to universities and to society at large. Data sharing, for example, will allow discoveries to be made more quickly in the areas of medicine and health, and environmental research into global warming by making research data open for all to use.

The pilot also needs to identify the skills, knowledge and training needed by all players in the data lifecycle, to suggest ways in which the requisite skills and knowledge can be gained and to determine who could provide training for the various stakeholders could be.

Infrastructure and Support Measures

Our organizations are well placed to provide support for projects under the data pilot. A solid support system is key to promoting and sustaining a healthy data-sharing ecosystem.

General information and support for project coordinators could be provided at various levels: through the European Commission and on disciplinary, European, national and institutional levels. In addition, our experience and access to broad research networks would greatly benefit an Open Data pilot.

OpenAIRE (www.openaire.eu) provides support and training to project coordinators and institutions which support researchers. This is done both through its Europe-wide network of National Open Access Desks (NOADs) and via services for data deposit and discovery through its e-infrastructure and data repository ZENODO (www.zenodo.org). Moreover, through harvesting repositories and mining techniques *OpenAIRE* infers links between publications, research funding and research data. This enables discovery and re-use, as well as monitoring and the ability to provide metrics.

LIBER (www.libereurope.eu) supports libraries in the development of institutional research data management policies and services. It also enables the exchange of experiences and good practices across Europe. Institutional infrastructures and support services are an emerging area and will be linked to national and international infrastructure and funder policies. Building capacities and skills, as well as creating a culture of incentives for collaboration on research data, management will be core targets of *LIBER*.

Together, the Confederation of Open Access Repositories (*COAR*, www.coar-repositories.org), *OpenAIRE* and *LIBER* are launching an Interest Group on the 'Long Tail of Research Data' within the framework of the Research Data Alliance (RDA). The group will focus on research data that falls outside of the 'big data' category. Much of the focus for improving the accessibility of data has to date been focused on big data: the massive datasets produced through large science projects. However, there are thousands of research projects every year that produce datasets. Many of these simply rest on the hard drives of researchers and are neither catalogued nor accessible. Attention will therefore focus on the data generated in universities and research institutions, and the role of institutional repositories and libraries as agents of the institutional data management.

For More Information

Please contact the following people for more information:

OpenAIRE: Yannis Ioannidis, OpenAIRE Coordinator, University of Athens, yannis@di.uoa.gr
Donatella Castelli, OpenAIRE Technical Coordinator, donatella.castelli@isti.cnr.it

LIBER: Paul Ayris, LIBER President, University College London, p.ayris@ucl.ac.uk

COAR: Norbert Lossau, Chair COAR Executive Board, OpenAIRE Scientific Coordinator, University of Goettingen, Norbert.Lossau@zvw.uni-goettingen.de