**LODES: A Study to Define a Linked**

**Open Data Platform of Interoperable**

**Repositories to Enable Open Science**

Scoping Group Report

R. Barbera, S. Bianco, M. Maggi,

D. Menasce, L. Patrizii - INFN - Italy

M. Saccone, L. Trufelli, L. Reggiani - CNR - Italy

R. Smareglia - INAF - Italy

P. Budroni - University of Vienna - Austria

+

All other Authors identified during LODES18 Workshop,

and willing to contribute to the report afterwards

L. Alagna, A. Masiero

Representatives of Italy to the OECD Global Science Forum

**DRAFT**

2018.08.24

Abstract

The main objective of this project is to carry out a study and deliver a report on enablers, obstacles, policies and recommendations for the setup of a cross-country federation of standard-based, multi-domain Linked Open Data repositories, leveraging as well as addressing the concept of Open Science.

# Introduction and context

# *(The present report originates from the proposal [1] and is meant to trigger the establishment of a Working Group whose activity would be the natural continuation/evolution of those that led to [2] and [14]. This is fully coherent with OECD [15] and GSF [16] missions).*

In its pioneering report *“Making Open Science a Reality”* [2], published in 2015, OECD defined Open Science as *“a means and not an end in itself and […] much more than just open access to publications […]”.* Among the various Open Science aspects and enablers discussed in that document, opening research data was considered key not only to accelerate scientific discovery and improve science reproducibility, but also to speed up innovation and improve citizen engagement with research, hence benefiting the whole society.

On October 18, 2017, an OECD-GSF “Workshop on Open Data” - hosted by INFN, CNR Ministry of Research with the participation of the Italian scientific community - discussed recent activities on Open Science repositories and concluded on the existence of two areas of interest:

* Quality and certification of Open Data: trust building, re-fitness or use and reusability (certification and reproducibility);
* Open Data repositories’ issues: multi/trans-disciplinary use of data, including interoperability and standards.

In order to address interoperability and federability of Open Data repositories, the concepts of FAIR (Findable, Accessible, Interoperable, Reproducible) data [3] and Tim Berners-Lee’s “star-based” classification of Open Data [4] have been coined and they are actively pursued at various political [5,6] and policy making levels [7-11]. However, shared recipes and widely accepted guidelines for the implementation as well as elastic and agnostic/transparent [12] growth of large scale, cross-country, cross-disciplinary and multi-purpose FAIR and “5 stars” compliant Linked Open Data (LOD) infrastructures are still missing.

The importance of Open Data was also extensively underlined throughout the book *“Opening Science”*[13]. Indeed, Open Data as a concept and Data Repositories (DRs) as the implementation of that concept, are indicated as key elements of the Democratic and the Infrastructure “schools of thought” of Open Science, respectively (see chapter *“Open Science: One Term, Five Schools of Thought”* in [13]).

Complexity, variability and heterogeneity are all defining attributes of modern Big Data Science and the concept of “openness”, in such a multi-dimensional data universe definitely requires the establishment of a common framework to insure its interoperability among all stakeholders and users. It is believed that such a goal can only be reached by adopting a shared, unitary approach taking into account both the large phase space of the problem along with the specificities of the individual communities involved.

# Stakeholders and technical practices

# *(Identify stakeholders (researchers, policymakers, ICT specialists, academic institutions, citizens, industries, etc. Identify and analyse a multiplicity of different cultural contexts, practices, approaches and the relative technical languages).*

# Open Data and Linked Open Data are generic terms that have different connotations among different scientific disciplines. Researchers in these domains have vastly different requirements, from biology, to particle physics, to medicine, geology, to humanities and social sciences. Identifying the stakeholders translates, in part, to pointing out the different needs and the correspondingly appropriate solutions for each community, building on a common set of rules and methodologies. In order to reach a consensus on this common set, which makes the Open Data and Linked Open Data approaches sustainable, both in implementing them as well as in maintaining them, policymakers need to be properly informed by scientists, ICT specialists and researchers in general, about what is required, what the possible solutions are, what are likely data, processing and network infrastructures providers and funders. Aim of the proposed scoping paper is exactly this process of accurate identification of the stakeholders, their requirements and an analysis of the possible technical solutions.

# 

# *A list of Open Data repositories and Linked Open Data services currently existing and willing to participate in a wider OECD Working Group will be identified during the LODES Workshop.*

# Challenges and Opportunities

*(Address several important questions of Open Data, such as: In the context of data handling, what are the lessons learnt from Big Science experiments that can be applied to “Long Tail of Science” groups? What are the technical/political challenges to achieve full interoperability among Open Science repositories? Who are the key actors for the deployment of Open Science repositories? What is the role of Data Stewardship? Who owns the data? Who is entitled to provide the certification of the quality of an Open Science repository? Is it possible to broadly categorize the type of information in Open Science repositories (Big-Science/Research, Citizen demographics, Social and Human Sciences, Law Studies, Cultural Heritage, Health, Environment, etc.)? Can the above disparate communities use a common infrastructure? What is the correct granularity needed to define metadata semantics to satisfy the different community requirements? What is the role of private companies and societal organisations?)*

Creating an Open Data infrastructure, capable of addressing the needs of scientists in a vast array of scientific domains is a challenge which extends across multiple dimensions. On one side there is a very large number of rather small communities, the so called “long tail of science” which own data scattered in innumerable data bases, under different and incompatible protocols, very difficult to access in a unified, standardized queuerable way. On the opposite extreme is a much smaller number of very big collaborations which own humongous data sets, like the four LHC experiments at the CERN laboratory, with data hosted in data centers scattered all around the world. In between is a continuous spectrum of data samples ranging from the small/medium size to the very large ones.

An important challenge to address is the level of accessibility of these data. In domains like biology and medicine, security and privacy are extremely important requirements, not necessarily so in physics. On the other hand, in the latter domain, data large volumes make it difficult to identify cost effective and efficient strategies for data accessibility.

Another key issue is to identify the actors for the deployment of such an infrastructure. On one side, ICT specialists, managers of large, distributed data-sets are certainly key experts, but equally so are the experts in the different scientific domains, namely researchers who do actually use these data to produce results. Merging the two communities, bringing them together to speak the same (meta)language being able to mutually understand the technical difficulties in the project is a challenge in its own.

Of fundamental importance is the issue of data ownership and stewardship. For each and every researcher and even more so for each Funding Agencies, data represent the very first output of their investment. Investigating and clarifying the subtle issues inherent in these concepts is another key point of the proposed scoping paper.

Finally, certification of the quality of the data provided by an Open Science repository will be addressed. Data are the base on which the scientific article finds its foundations. The problem of certifying the quality of a scientific article has been tackled differently depending on the field of research. For science sectors such as physics, mathematics, chemistry, etc, the quality certification proceeds mainly via a peer-review process handled by major commercial publishers, while other fields may have different approaches, with very relevant implications on the controversial market of scientific publishing. How to enlarge the certification process to Open Data is a critical issue with far-reaching implications: market economy, research infrastructure expenditure, science evaluation criteria, funds assignments, career progressions.

# Standards, best practices and guidelines

(*Aim of the LODES workshop is to collect requirements and needs from the identified stakeholders and then propose the creation of an OECD Working Group charged with the definition of standards, best practices and guidelines for a transnational science policy aimed to the definition and development of a LOD federation of repositories*).

The vast range of scientific communities, together with the large number of possible data access modes makes it rather difficult to identify a simple set of standards: what is advisable, at this stage, is a preventive, in-depth, survey of both needed requirements and already existing standards, from which a comprehensive set of solutions could be built upon. Aim of the scoping paper is exactly to explore all aspects of this assessment activity.

# Possible demonstrative use cases

*(A few use cases, which will be identified during the LODES Workshop and could act as demonstrators in the wider context of an OECD Working Group, will be briefly described in this section).*

# Conclusions and Outlook

*(Conclusions will be drawn at the end of the workshop, with whatever consensus will be reached on the goals, means and methodologies to setup a concrete activity aimed at providing our communities of the needed policy tools for an efficient, interoperable Open Data federation. The outlook, for now, is to be successful in creating an active community around this challenging idea*. *On the Working Group timeline, if approved, it will be of the utmost importance to identify how to implement these concepts in a fully fledged report, optimising existent funds and new dedicated resources. The kind of expertise - such as information technology, semantic web, librarian, etc. - required by those in charge of preparing the final Report of the Working Group is an important issue to discuss, in order to provide feedback to the Funding Agencies).*

# References

# R. Barbera et al., A Study to Define a Linked Open Data Platform of Interoperable Repositories to Enable Open Science [PUBLICATION INFN-2018-001] 10.15161/oar.it/1525110414.75

# OECD, “Making Open Science a Reality” (<http://dx.doi.org/10.1787/5jrs2f963zs1-en>).

# <http://www.force11.org/group/fairgroup/fairprinciples>

# <http://5stardata.info>, Tim Berners-Lee’s 5-star deployment scheme for Open Data

# European Commission, “Open innovation, open science, open to the world - a vision for Europe” (<http://dx.doi.org/10.2777/061652>).

# European Open Science Cloud (<https://goo.gl/gzYpNo>)

# High Level Expert Group on the European Open Science Cloud (<https://goo.gl/hEt61N>)

# Workshops of European Open Science Cloud (<https://goo.gl/97TYoK>)

# European Commission, “Guidelines on FAIR Data Management in Horizon 2020” (<https://goo.gl/MwiHqX>)

# EOSCpilot Project, “D3.1: Policy Landscape Review” (<https://eoscpilot.eu/sites/default/files/eoscpilot-d3.1.pdf>)

# EOSCpilot Project, “D6.3: 1st Report on Data Interoperability: Findability and Interoperability” (<https://eoscpilot.eu/sites/default/files/eoscpilot-d6.3.pdf>)

# The term “agnostic/transparent” is used with the meaning explained here: [https://en.wikipedia.org/wiki/Transparency\_(human%E2%80%93computer\_interaction](https://en.wikipedia.org/wiki/Transparency_(human%E2%80%93computer_interaction))

# “Opening Science - The Evolving Guide on How the Web is Changing Research, Collaboration and Scholarly Publishing”, Sönke Bartling and Sascha Friesike (Eds.), Springer 2014 ([http://book.openingscience.org](http://book.openingscience.org/))

# OECD, “Business models for sustainable research data repositories” (<https://goo.gl/AX8e69>)

# <http://www.oecd.org/about/>

# “THE OECD GLOBAL SCIENCE FORUM - Strategic Directions 2015-2019” (https://www.oecd.org/sti/sci-tech/STI-GSF-brochure.pdf)

# 