e-Infrastructures: The starting blocks for Open Science and Innovation

The OpenAIRE case

Natalia Manola

Research Associate, Athena Research and Innovation Centre
Managing director of OpenAIRE

Abstract. There is an emerging consensus among the members of the academic research community that eScience practices should be congruent with Open Science, which essentially facilitates multi-disciplinary collaboration and data sharing and re-use, entailing Open and Shared Access to research data, data services, tools, analyses, and methods. This short paper presents the key characteristics of **a sustainable open science e-Infrastructure** as gained from six years of operation of OpenAIRE, a key EU e-Infrastructure on scholarly communication. presented.

Keywords: open science, infrastructures, open access policies, text and data mining licenses

1 About Open Science

Open Science has different meanings to different people, but the emergent principles consistently emphasize openness, sharing, and democratization of all scholarly and research resources and outcomes. In general, we're moving towards an era of greater transparency in all processes related to methodology, data, communication, and collaboration, and **e-Infrastructural services are key in shifting processes and behaviours towards the right direction**.

The problems we face in gaining widespread support for Open Science are really about incentives and sustainability. The ultimate question we need to respond to is *"How can our efforts affect or modify the scientific reward systems to make Open Science the natural state of affairs for researchers?"* Right now, researchers mostly favour "closed" science, with clear disincentives towards openness. Researches are people, and are motivated by most of the same things as normal people[1]:

- Money, for themselves, for their groups, and to support their science.

[1] Dan Gezelter, "What is exactly Open Science?"
Blog http://www. Openscience.org/blog/?p=269

- Reputation, which is usually (but not necessarily) measured by citations, h-indices, download counts, placement of students, etc.
- Sufficient time, space, and resources to think and do their research (which is, in many ways, the most powerful motivator).

E-Infrastructures, and particularly OpenAIRE, work on lifting the existing disincentives, addressing the reputation and resource efficiency/optimization aspects, aiming to lower the barriers for engaging the different stakeholders.

2 What is OpenAIRE

OpenAIRE is a socio-technical network that supports the *implementation* and *monitoring* of Open Science policies, including Open Access to publications and research data:

- **Implementation** is enabled by a pan-European network of Open Access/Open Science experts – the National Open Access Desks (NOADs)[2], present in every EU country and beyond. The NOADs work together to align national policies, define shared solutions and best practices, and coordinate outreach and advocacy activities through a range of targeted training events and support materials.
- **Monitoring** is achieved by means of an advanced data infrastructure consisting of a decentralized network of data sources, namely publication repositories, data repositories, and current research information systems, established by research institutions, individual scientific communities, and publishers. By harnessing the contents of "compatible" publication, data, software, and method repositories (both institutional and disciplinary) and linking them to other research entities (researchers, institutions, projects), OpenAIRE produces a 360˚ picture of the impact of European research funding.

3 Key characteristics of a sustainable open science e-Infrastructure

3.1 One size does not fit all – Coordinate a participatory and decentralized design

Open Science needs pragmatic, participatory infrastructures to work. The diverse and culturally varied research communities of Europe will **accept no "one-size-fits-all" solutions**. While some see variety and diversity as insurmountable barriers, for OpenAIRE, they are the foundation stones upon which we build. Embedded in a global network of **regional repository networks**, OpenAIRE is a component in the global scene, considered an exemplar for other regions in the world, placing Europe at the forefront of Open Science developments.

[2] https://www.openaire.eu/contact-noads

The need to incorporate a diverse network of repositories (publication, data, software) and to approach and embed different disciplines, needs and behaviours in this network, effectively means that **interoperability** must take place at all levels. Even though we usually focus the need on the technological interoperability, it is crucial to find ways to interoperate on the legal, policy and cultural levels, which is only achieved through a **participatory** design.

E-Infrastructures have placed considerable effort in decentralization, i.e., the decoupling of the underlying storage and access services from the value added services, a trend which is now being applied on the Web[3], and are paving the way to an open and transparent environment that results in true data ownership and improved privacy, essential elements in a trusted open science environment.

3.2 Open scholarship is in a flux - Build adaptable and trusted systems

In the past decades, technological advances in the areas of distributed computing, high-performance computing, data management, and networking have changed the paradigm of research endeavours. Scientists and scholars from many disciplines are now able to conduct large-scale experiments or observations, analyse tremendous amounts of data generated by these, draw significant scientific conclusions based on the results of such analyses, and present these results in traditional or electronic publications. Individual researchers, members of a laboratory, collaborators in the context of a project, or whole scientific communities are storing the data and publications that are relevant to their research, which if taken all together create an extremely rich global resource that is invaluable for further and future research work.

A number of emerging changes affect the way science is conducted and communicated, which require systems that **are able to adapt to changes fast**, are **not bound to specific technologies or business models** (avoid lock-in mechanisms), include **openness and transparency** at all levels to foster trusted, global uptake:

- **Data driven science and shared economies** call for the ability and the need to share facilities and data.
- **Reproducibility, replicability and accountability** in science plays a crucial role as results drive decision making processes that have a global effect.
- **Scientific and scholarly credit goes beyond peers** and reaches out to broader audiences with socioeconomic effects, affecting the overall research assessment.
- The social web has spurred **new types of collaborative behaviours** amongst researchers, who use collaborative tools and services[4] even from the early processes of the research life cycle.

[3] https://solid.mit.edu/

[4] e.g., Authorea, Winnower, Jupyter Notebook

3.3 Research is global, support is local – Invest in building and sustaining human capacities

Research is a globalizing endeavour. Whether this takes place in collaborative global environments assisted by research infrastructures, or it happens individually in research organizations, the so called long tail of science, research does not stay anymore within the previously established discipline or regional borders. An important part of open scholarship is a barrier-free communication, which at present happens not only at the final publishing stages, but throughout the research life cycle. These eScience/open science new practices require advanced knowledge and skills, embedded strategically and supported continuously in the research environment. **Modern e-Infrastructures need to take into account this building and sustaining of human capacities as a key success factor for achieving their goals.**

In particular, data skills are important in all phases of the research process, and in the case of publishing, where OpenAIRE comes into play, the libraries play a key role in supporting and improving these practices. The new generation of "data librarians" needs to be educated so as to guide and train researchers on a variety of topics, including the following:

- availability and use of tools and services from commercial or public e-Infrastructures (whether these are national/EU/global thematic or more generic e-Infrastructures) to be embedded into the library processes;
- promotion and use of open tools and their effective integration into the research processes;
- different modes and support of publishing for articles and data (electronic publishing) and the accompanying business models (costs and benefits that go beyond the financing aspect);
- support of data curation and stewardship and relevant best practices.

OpenAIRE has created and operates an **active human network** with representation from 33 EU+ member states, the National Open Access Desks (NOADs). These are experts familiar with national and local developments in their specific jurisdiction, cultural, governance and financing systems, providing **support and solutions** for policy implementations, aligning and transferring **practices and new ideas** for research or publishing practices. With their continuous engagement with the research community they are key in guiding the transition to and long-term implementation of Open Access and Open Science policies.

3.4 Open Science services for all – Move towards Open Science as a Service (OSaaS)

Key to the uptake of open science practices by researchers is the **availability of trusted, secure and robust value added services** that are embedded in all phases of the research. All stakeholders in the chain of science, i.e., funding agencies, organizations, researchers, are demanding researchers to marry *Open Science* paradigms in order to ensure that all kinds of digital artefacts (literature, research data, research

methods, and others) of a research activity can be (as openly as possible) shared, possibly in relation with their experimental context. The ultimate goal is to offer researchers all the components and services they require to *repeat* ("same experiment, same laboratory"), *replicate* ("same experiment, different laboratory"), *reproduce* ("same experiment, different input parameters"), or *re-use* ("using a sub-part of the experiment into another experiment") the research activity.

E-Infrastructures support the implementation of such services, especially as they by design decouple the low/soft services from the end user services. The latter frequently materialize with the help of 3rd party providers (public or commercial), who are able to skip the mundane, time consuming steps of reaching to multiple stakeholders, of homogenizing and maintaining the data used in these services.

Furthermore, the recent Open Science as a Service (OSaaS) approach[5] uses the horizontal e-Infrastructure services to deliver out-of-the-box, on-demand deployable tools in support of Open Science. An innovative concept by itself, it effectively allows research communities, content providers and other stakeholders to select from a **portfolio of services** built on top of horizontal e-Infrastructures, to configure and adapt them for their specific domain. In addition, it brings a change in the way researchers use, uptake, familiarize themselves with, and ultimately become an integral part of a global interconnected and seamless e-Infrastructure ecosystem.

But researchers are not the only group in need of reliable services. Funders, research managers and other decision makers rely on open, trusted and transparent data so as to better connect to and understand the research processes, to evaluate its results to support evidence-based decision making, to monitor policies (beyond open science) and their effectiveness, to measure the broader socioeconomic impact, to identify research trends for future funding. These services are also in accord to the Open Science principles and are best served by horizontal e-Infrastructures.

From its onset, OpenAIRE has pursued a **service-driven design** to engage all stakeholders, to change mentalities and behaviours, and to ensure the uptake of OA. Its service portfolio provides a rich set of services, implemented at different e-Infrastructure layers (repositories, mediation platform, text and data mining, value added services) targeting a variety of stakeholders (funders, researchers, data providers, SMEs, academic networks, the public), demonstrating with the continuously increasing uptake of its services the central role of e-Infrastructures in the Open Science scene.

3.5 Community driven and publically funded – Involve the research community in the design, operation and governance

Community driven e-infrastructures (i.e., designed, governed and operated) reveal real researchers needs more intimately than commercially driven e-Infrastructures or services, as the current situation with publishers shows. This ensures that i) the design reflects researcher trends and behaviours in an agile way, and ii) there are no lock-in solutions in technologies, legal, or business models which inherently prohibit future

[5] http://www.slideshare.net/OpenAIRE_eu/open-science-publishing-asaservice

improvements. Engaging the wider community via representatives, e-infrastructures can easily adapt to new trends, adopt new services as they come along, and more importantly, help invent new scholarly communication mechanisms. In addition, a publicly funded and publicly governed e-Infrastructure will guarantee persistence and sustainability, and ensure that outcomes are driven by scientific excellence and societal needs rather than profit. [6]

OpenAIRE, through its network of the National Open Access Desks who act as ambassadors to all stakeholders and the connection to research libraries who have a direct communication and contact with researchers, is entirely community driven. This has already built an inherent trust as it ensures that the **research community is in charge of future developments in open scholarship**.

3.6 Build trust for all stakeholders – Focus on quality, credibility and sustainability

All researcher user groups, no matter how diverse, choose services and infrastructures based on their credibility. Knowledge that the e-Infrastructures will be present for the long-term and that their services can be relied and built upon without fear for their future, builds trust amongst stakeholders and increases uptake of the e-Infrastructure.

The operation of OpenAIRE and our continuous interactions with our stakeholders has clearly shown that the following factors play a significant role in establishing trust in the community:

- **Community values** – respect and reflect the community approach. Do not enforce solutions, but stir decisions when possible.
- **Reliable services** – provide 24x7 trusted and secure services which can be easily embedded and integrated in researcher workflows.
- **Quality of data and content** – invest in data quality assessment processes at all levels of the e-Infrastructure (from the metadata and data stewardship at repositories or journals, to the final presentation of the aggregated content).
- **Sustainability** – ensure a robust business model that will allow services to persist over time.
- **Supporting services** – develop and operate supporting services (e.g., helpdesk, support, training) that will provide a better understanding for policies, implementation aspects, roles.
- **International recognition** – assume international leadership in the domain and support other regions to come up with interoperable solutions.
- **Service Level Agreements (SLAs)** – establish and follow business processes to attract 3rd party service providers.

[6] Position Paper: European Open Science Cloud for Research, https://zenodo.org/record/32915

3.7 Innovation is everywhere – Ease barriers to connect to the market

Open Science is presently considered by funders and ministries as a means to optimize cost of science and leverage innovation. E-Infrastructures are in the centre of the knowledge triangle (research, education, innovation) as their main objective is to nurture an open and transparent environment that cuts across disciplines and geographical borders, produces new knowledge and provides the opportunities for the creation of value added and innovative service and products. As such, e-Infrastructures are expected to have an immediate and a long term impact (i) in the production and acquiring of new knowledge for research communities that will be exposed to Open Science (new content, methods), and (ii) in the attraction of new businesses who, taking advantage of the enriched and interlinked open data, as provided by OpenAIRE, produce innovative tools/products (similar to the open government open data initiatives).

However, the principles of open policies can be widely accepted only if realized within a shared **integrated legal interoperability framework**, where the different legal rights, terms and conditions are harmonized and synchronized. The role of the e-Infrastructures is crucial in delivering such a framework, as this must involve a broad range of stakeholders, must be agreed upon and equally regulated by the research community and the public to ensure the right balance, as often (over)regulation stunts growth (e.g., lifting the Text and Data Mining licenses on digital and often paid for content).

More specifically, publications, or any narrative part of research, hide a wealth of information which can be brought out by text and data mining, and emerging AI processes. OpenAIRE is itself employing a set of these techniques to extract knowledge to use in internal workflows, but efforts like SemanticScholar, AMiner, Linknovate, Wheezbee, UberResearch, IRIS AI, Research[7] clearly show that possibilities of **reusing, combining and repurposing science** to new users are endless. As EC's project OpenMinted[8] is starting to demonstrate there is an active research community and SMEs waiting around the corner for a sustainable and trusted partner to get their hands on open research data to produce new knowledge, new services.

3.8 Policies hand in hand with technologies – Implement, monitor, assess

Policies – in order to be effective – **must go hand-in-hand with the infrastructures that support them.** Implementation is key to a successful policy and technical solutions must be in place *before* a policy is activated. Due to the diversity and different adoptions of solutions and requirements, the focus is on **interoperability** and accompanying **guidelines**, which are issued either at the data/metadata level or at the service level.

---

[7] https://www.semanticscholar.org/ , https://aminer.org/, https://www.linknovate.com, http://www.wheesbee.eu/, http://www.uberresearch.com/, https://iris.ai, http://www.researchresearch.com

[8] Text and data Mining e-Infrastructure for Europe, www.openminted.eu

Policies once issued are not, and should not be, static. They need to be **closely monitored** over a wide spectrum of disciplines and borders, to be **evaluated** with the present facts and to be **re-adjusted** when needed. The horizontal operation of e-Infrastructures ensures the proper monitoring mechanisms in a transparent, systematic, and reliable way as they span over a broad spectrum of research processes and outcomes and are then able to correlate with policies and provide **off-the shelf observatory** functionalities.

OpenAIRE in particular, enables the **linking of the whole spectrum of research outputs**, including articles, books, data, software, and other types of research outcomes. By attaching provenance information related to funders, organizations, people, facilities used, and by extending the links to all research artefacts (workflows, methods, protocols), it creates an interlinked database that captures, and subsequently monitors or recreates, all phases of the research life-cycle. OpenAIRE's production operation is comprised of a set of **big data processes** (aggregating, cleaning, transforming, de-duplicating, linking, extracting and inferring knowledge) that produce **curated, validated and interlinked research information.** This allows for the development of **off-the-shelf services policy monitoring** using a **shared economy** approach, e.g., an Open Science monitor to be used by different funders or institutions.

3.9 No investment wasted - Build on existing national infrastructures

Many of the resources and services needed for the implementation and support of Open Science already exist. Particularly in Europe, member states have already invested, or are in the process of investing in national research infrastructures (as published in the 2014-16 member states national roadmaps). These investments are taking into account open science policies and practices, and are looking into ways to link the research to the innovation process to boost the local research environment and market. In parallel, research performing organizations around the world have put in place the necessary resources to provide localized support and implementation.

What is still missing is a **coordinated effort** to bridge this decentralized but often fragmented ecosystem in order to ensure a better uptake and convince policy makers for **optimal public shared investments**. E-Infrastructures, with their horizontal services and their outreach to wide audiences, are the most fitting candidates to fill in these gaps as they align the different technical solutions, ease the barriers for policy, funding, interoperability, access policies, and provide a federated knowledge transfer. This coordinated and integrated service provisioning requires well-defined **rules of conduct** for all stakeholders involved, and will definitely support the transformation towards better and more open science and scholarship across Europe and beyond.

OpenAIRE, already using this model, relies on the existing **repository infrastructure, leveraging** the **existing investments** of institutional, thematic repositories and OA journals (both literature and data). All key stakeholders (researchers, libraries, administrators, funders, ministries) are involved in this process, showing that ***de facto*** **member states already share responsibilities, coordination and costs**.

3.10 e-Infrastructures are hard to grasp and commit to – Align and coordinate efforts

E-Infrastructures are multi purpose, multi faceted endeavours reaching out to a wide variety of stakeholders. Their role is frequently hard to grasp, as they provide several soft and/or low level services that are not well understood by all users. Researchers usually pay attention when value added services embedded in their daily activities go amiss, while funders lack the ability to see immediate return on their investments and often approach their sustainability with a cost rather than an investment mentality.

As with all other infrastructures, it takes time for researchers to change behaviours and practices and allow for the value of e-infrastructures to be noticeable, particularly via advanced services that are in the critical path of research. Therefore, they require continuous investment in people and money and long-term commitment from funders around the world (research is a global endeavour and e-Infrastructures only succeed if they address the challenges that arise at this level). Naturally, in doing the above, another layer of complexity arises, as these global efforts require alignment.

In this sense, e-Infrastructures require an effective coordinated effort that overcomes these barriers and brings to the same level of understanding people from different regional backgrounds and technical expertise. Such coordination may be best left to appropriate national organizations (e.g., national data services), with a regional or global overarching alignment.

4 What comes next?

Open Science is here to stay. The issues we now need to address are about what it really means and how it is translated to specific implementation steps. To be successful, we need to set the right priorities, to identify and assign the appropriate roles and responsibilities on the service provision side, and to come up with well defined rules of conduct for all participating bodies and end users.

EU's vision for an European Open Science Cloud (EOSC), as this is expressed in the common vision paper from EUDAT, LIBER, OpenAIRE, EGI, GEANT[9], and the EOSC High level Expert Group recommendations[10] sets some of the open science principles and priorities in motion, and places e-Infrastructures in the forefront. OpenAIRE, as it has already established itself as a key infrastructure for Open Science in Europe with a well recognized brand name, takes a first and decisive step in the implementation of the EOSC.

[9] Position Paper: European Open Science Cloud for Research, https://zenodo.org/record/32915

[10] A Cloud on the 2020 Horizon. Realising the European Open Science Cloud: first report and recommendations, http://ec.europa.eu/research/openscience/pdf/hleg/hleg-eosc-first-report_(draft).pdf#view=fit&pagemode=none