The emergence of open data

David Ball, David Ball Consulting

davidball1611@gmail.com

Summary
Open Science is moving centre-stage, with a rationale of improving efficiency in science; increasing transparency and quality in the research validation process; speeding the transfer of knowledge; increasing knowledge spill-overs to the economy; addressing global challenges more effectively; and promoting citizens' engagement in science and research. Open Data has undergone a surge in practical development, mirroring the well-established repositories for research outputs. The development and application of model policies and of principles are discussed and the views of researchers championing Open Data highlighted.

1. Open Science

1.1. Why Open Science?

The concept of Open Access (OA) to research outputs such as journal articles has been common currency for many years. The seminal Budapest Open Access Initiative (BOAI) statement of February 2002, for instance, reads:

> *"By open access to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited."*

More recent thinking, however, for instance by the European Commission, has expanded the concept of openness even further, to Open Science (OS), which aims to transform science by making research more open, global, collaborative, creative and closer to society. This shift is potentially very important for the development and exploitation of research.

OS is about the way research is carried out, disseminated, deployed and transformed by digital tools, networks and media. It relies on the combined effects of technological development and of cultural change in the direction of collaboration and openness in research.

To elaborate, an OECD report (1) identifies the following six rationales for policies that seek to implement and support OS, including Open Data:

- *Improving efficiency in science* – OS can increase the effectiveness and productivity of the research system, by: reducing duplication and the costs of creating, transferring and re-using data; enabling more research on the same data; multiplying opportunities for domestic and global participation in the research process.
- *Increasing transparency and quality* in the research validation process, by allowing greater replication and validation of scientific results.
- *Speeding the transfer of knowledge* – OS can reduce delays in the re-use of the results of scientific research, including articles and data sets, and promote swifter development from research to innovation.
- *Increasing knowledge spill overs to the economy* – Increased access to the results of publicly funded research can foster spill overs and boost innovation across the economy as well as increase awareness and conscious choices among consumers.
- *Addressing global challenges more effectively* – Global challenges require co-ordinated international actions. OS and Open Data can promote collaborative efforts and faster knowledge transfer for a better understanding of challenges such as climate change, and could help identify solutions.
- *Promoting citizens' engagement in science and research* – OS and Open Data initiatives may promote awareness and trust in science among citizens. In some cases, greater citizen engagement may lead to active participation in scientific experiments and data collection.

1.2 What it is

Each step of the research lifecycle is becoming more open, for instance through:

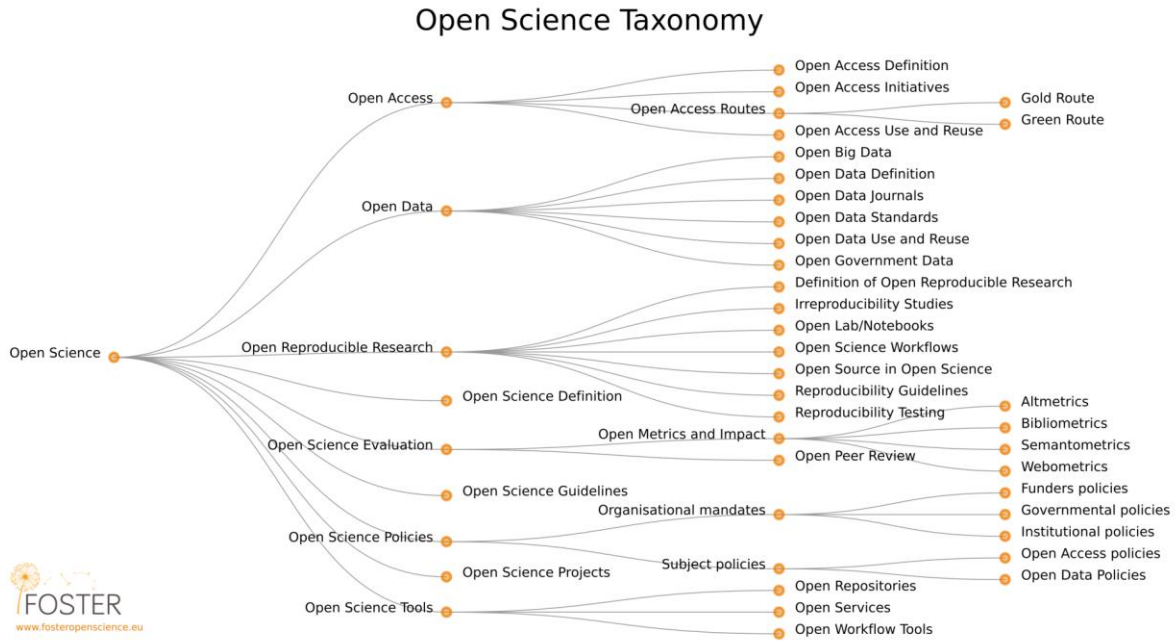
Open Notebooks - an emerging practice, documenting and sharing the experimental process of trial and error;
Open Data - managing research data in a way that optimises access, discoverability and sharing for use and re-use;
Open Research Software - documenting research code and routines, and making them freely accessible and available for collaboration;
Open Access - making all published outputs freely accessible for maximum use and impact.

In order to achieve this openness in science, each element of the research process should:

- **Be publicly available** - it is difficult to benefit and use knowledge hidden behind username and password barriers, or if it does not contain the right metadata to make it discoverable.

- **Be re-usable** - research outputs must be licensed appropriately so that prospective users know clearly any limitations on re-use.
- **Induce collaboration** between researchers through better access and better online tools;
- **Be transparent and have appropriate metadata** to provide clear statements of how research output was produced, and can be re-used.

A more concrete exposition of Open Science and its many branches is provided by the taxonomy developed by the European FP7 FOSTER project in support of its aim of putting in place "sustainable mechanisms for EU researchers to FOSTER OPEN SCIENCE in their daily workflow, thus supporting researchers optimizing their research visibility and impact, the adoption of EU open access policies." The taxonomy covers not only the constituent elements of OS but also supporting tools, measurements and mechanisms.



1.3 How open is our research?

Further evidence of this widening from Open Access and Open Data to Open Science is provided by the development by SPARC Europe of a tool for visualising, discussing and monitoring how open an institution's research is. It takes the form of a radar diagram generated by confirming status or actions, or estimating percentages, in eleven main topic areas:

The eleven topic areas are exhaustive and demonstrate the potential extent and complexity of Open Science, and the challenges faced by institutions and funders in bringing it about.

Such diagrams can be used in a number of ways, for instance as an assessment tool, to generate discussion and to inform policy-making.

2. Open Data

2.1 What is it?

As we noted at the start of this document, while Open Access to research outputs has a long history and development, Open Data has come into scope somewhat later. The OECD report already quoted makes the rationale specific (2):

> *"… reducing duplication and the costs of creating, transferring and re-using data; enabling more research on the same data; … increasing transparency and quality in the research validation process, by allowing greater replication and validation of scientific results."*

Research data can be defined simply as whatever is either produced in research or evidences research outputs.

The European Commission's definition is: "information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation". Examples include: statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings, images (3).

The 2012 European Commission's recommendation on access to and preservation of scientific information states that: "open access to scientific research data enhances data quality, reduces the need for duplication of research, speeds up scientific progress and helps to combat scientific fraud" (4). Unsurprisingly other funders also require open access to the data produced as a result of their funding. The Wellcome Trust for instance has been a leader in the field. Its *Policy on data management and sharing* (2010) states: "The Wellcome Trust expects all of its funded researchers to maximise the availability of research data with as few restrictions as possible."

Over time policies have developed. Commonly they will now include the following elements (5):

- **Timing**: when publication should take place;
- **Data plan**: requirements for a technical management plan;
- **Access and sharing**: what exactly will need to be available for public use;
- **Long term curation**: data creation and sustainability;
- **Monitoring**: any monitoring that will be carried out by the funding body and guidance available;
- **Storage**: details of the appropriate repository or data centre to be used;
- **Costs**: where costs can be claimed from and when.

Making data open is in some, particularly technical, senses more complex than making research outputs open: data collected must be capable of being verified, processed and re-used. However there are many resources covering all aspects of Open Data policies and practice now made available, for instance, by the Digital Curation Centre.

HEFCE, Research Councils UK (RCUK), Universities UK (UUK) and Wellcome recently (July 2016) published *Concordat On Open Research Data*, which is meant to:

" … [help] ensure that the research data gathered and generated by members of the UK research community is made openly available for use by others wherever possible in a manner consistent with relevant legal, ethical and regulatory frameworks and norms … The intention [of the Concordat] is to establish sound principles which respect the needs of all parties. It is not the intention to mandate, codify or require specific activities, but to establish a set of expectations of good practice with the intention of establishing open research data as the desired position for publicly-funded research over the long-term."

The development and promulgation of such principles is welcome. The *Concordat* also gives clear definitions and examples demonstrating that data are the result of humanities as well as scientific research:

"Research data are the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence… They may include, for example, statistics, collections of digital images, sound recordings,

transcripts of interviews, survey data and fieldwork observations with appropriate annotations, an interpretation, an artwork, archives, found objects, published texts or a manuscript."

The *Concordat* also recognises that:

"Not all research data can be open and the Concordat recognises that access may need to be managed in order to maintain confidentiality, guard against unreasonable cost, protect individuals' privacy, respect consent terms, as well as managing security or other risks."

In its *Guidelines* cited above, the European Commission takes as its Open Data mantra "as open as possible, as closed as necessary" and gives more detailed exemptions to data whose publication would be:

- Incompatible with the obligation to protect results that can reasonably be expected to be commercially or industrially exploited;
- Incompatible with the need for confidentiality in connection with security issues;
- Incompatible with rules on protecting personal data.

FORCE11 has also published the FAIR data principles. Data should be:

- **Findable**: easy to find the data and the metadata for both humans and computers - persistent identifiers (PIDs);
- **Accessible**: data should be retrievable by their identifier using a standardised and open communications protocol;
- **Interoperable**: data should be able to be combined with and used with other data or tools. The format of the data should therefore be open and interpretable for various tools;
- **Re-usable**: metadata and data should be well described so that they can be replicated and/or combined in different settings.

2.2 What do researchers think of it?

Open Data may be in its infancy, but already there are outspoken champions among researchers. The views of fourteen researchers from seven different countries, active in diverse disciplines, were collected and published in 2017 by SPARC Europe. A number of themes emerge.

There are various prominent **rationales** for Open Data:

- Data produced as a result of public funding should be publicly available.
- It is only possible to validate or reproduce research findings if the underlying data and tools are available. Otherwise they have to be taken on trust.

- As is now very often the case, independent research groups around the world creating their own data gives rise to inefficiencies.
- Data can often be re-used, for instance being subject to different methodologies or coupled with other data.
- Open Data alone makes possible the creation of very large data sets, which can be exploited by machine techniques such as data mining.

The following actions are recommended to foster Open Data:

- There needs to be a change in research culture, so that sharing data becomes the norm. This of course depends on the motivation of researchers through academic incentives.
- Such cultural change may be speeded by gaining the active support of senior researchers and managers.
- Funders' policies can play a very significant role in achieving such change.
- Open Data (and Open Science) must become an integral part of researchers' education, not something separate.
- It must be made as easy as possible for researchers to deposit and share their data.

Some characteristics of an **Open Data world** are identified as:

- Knowledge creation will be accelerated, producing real-world benefits, particularly for medicine and business.
- It should be possible to draw on or incorporate large data sets created outside academia, for instance in transport, meteorology and medicine.
- The ready availability of data, with appropriate metadata, should drive the development of interdisciplinary research.

A very few downsides to Open Data were identified, including possible breaches of confidentiality and researchers' perception of the data they create as being their own property.

3. The Future

Open Science is moving centre-stage, with a rationale of improving efficiency in science and speeding the transfer of knowledge. We have seen a surge in practical developments for Open Data, mirroring the well-established repositories for research outputs, and in the development and application of model policies and principles.

It is very easy to become blinded by the interesting detail of these advances. However it is salutary to paraphrase Bill Clinton's mantra on the economy: "it's the research, stupid". It is researchers themselves, funders, governments, supra-national bodies such as the European Commission and industry and commerce that will benefit directly from and drive this openness. The benefits are potentially, huge, multiplying the return on investment in research, accelerating the research process and involving a full range of interested citizens.

We can already see new paradigms and structures arising. As information professionals we are already closely involved in their development. We must be seen to be leading the move to the new pervasive openness.

References

(1) OECD 2015.
(2) OECD 2015. Making Open Science a Reality.
(3) Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020
(4) European Commission 2012 recommendation on access to and preservation of scientific information.
(5) Guy, M. and Ploeger, L. 2015. PASTEUR4OA Briefing Paper: Open Access to Research Data.