Open Science
Practical Issues in Open Research Data

Robert Viseur[1,2]
[1]*CETIC, Rue des Frères Wright, 29/3, 6041 Charleroi, Belgium*
[2]*Faculty of Engineering, University of Mons, Rue de Houdain, 9, 7000 Mons, Belgium*

Keywords: Open Science, Science 2.0, Open Innovation, Open Source, Open Data, Open Research Data.

Abstract: The term "open data" refers to information that has been made technically and legally available for reuse. In our research, we focus on the particular case of open research data. We conducted a literature review in order to determine what are the motivations to release open research data and what are the issues related to the development of open research data. Our research allowed to identify seven motivations for researchers to open research data and discuss seven issues. The paper highlights the lack of dedicated data infrastructure and the need for developing the researcher's ability to animate online communities.

1 INTRODUCTION

The word "open data" refers to "*information that has been made technically and legally available for reuse*" (Lindman and Tammisto, 2011). Open data draws interest, because of: the developments in scientific research (concept of reproducible research and sharing of experimental data); the enthusiasm, especially within the scientific community, for the semantic Web and the linked data; the publications of datasets in the public sector (e.g. geographical information); and the emergence of online communities (e.g. OpenStreetMap). The open data movement engages the public sector, the businesses and the researchers. In the scientific community, the open data adoption is linked to the recent evolutions in scientific practices. The latter include the online data sharing and the online sharing of unfinished works that allow to accelerate the process of discovery and get feedbacks about the research (Teif, 2013).

Our paper is dedicated to a particular case of open data: the open research data. The open research data falls within the context of so-called science 2.0 and open science. The first one refers to the use of Web 2.0 practices and tools in the field of scientific activities. The second one refers to openness. Practically it can be linked to open innovation and open source that are popular since a decade. Our paper is organized in three sections. In the first section we present the motivations for opening the research data. In the second section we offer an inventory of the issues related to the opening of research data and propose some solutions to address the issues. In the third section we resume our findings and discuss the perspectives of our research.

2 MOTIVATIONS

Some motivations explain the interest in the scientific community for the open research data.

2.1 Open Access

The open access for the scientific and scholarly research texts, as defined in the Budapest Open Access Initiative (www.budapestopenaccess initiative.org), means *"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 open access gains popularity through true open access journals (e.g., PLOS One) and preprint archives (e.g., ArXiv.org) allowing to deposit the papers after an embargo period.

Viseur R..
Open Science - Practical Issues in Open Research Data.
DOI: 10.5220/0005558802010206
In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (DATA-2015), pages 201-206
ISBN: 978-989-758-103-8

Unfortunately, the words "open" and "open access" stay to be used with various meanings and may cause confusion (Murray-Rust, 2008). The publication of research data with the paper is in line with the objectives of open access. It is encouraged by subsequent initiatives such as Panton Principles for Open Data in Science (pantonprinciples.org).

2.2 Reproducible Research

In computational research the verifiability of published findings needs the access to data and computer code. That observation conducts to the concept of "reproducible research" that would belong to Claerbout (1992). The latter highlighted that one of principal goals of scientific publications was to *"provide enough detail to make the work reproducible*". Many years after his observation that "*in real life, reproducibility is haphazard and variable*", progress has been slow. For example, Vandewalle et al., (2009) recently detected failures of reproducibility in the papers they analyzed in signal processing field. Their study is based on an analysis grid to assess the current reproducibility practices and defines six degrees of reproducibility.

Thus the journal Insight (www.insight-journal.org) shows how open access, open source and open data could change scientific practices. It is an online publication with peer reviewing that is associated with the software Insight Segmentation and Registration Toolkit (ITK). The latter is supported by the company Kitware (www.kitware.com). ITK is an open source software tool for image analysis (www.itk.org). The scientific results are published with the article (as traditionally) but also with the source code and the data in order to enhance the reproducibility of the researches ("reproducible researches") (Jomier et al., 2007). The newspaper technical infrastructure automates the source code compilation and testing.

Other projects such as GNU Octave highlight those practices (e.g., Eaton, 2012). Several authors went one step further and developed the concept of "executable papers" based on the linked data and the cloud infrastructures (Kauppinen and Espindola, 2011). In addition to publish papers with data and code, the idea is to provide virtual machines ready to execute on the cloud or on a local computer without dealing with dependencies issues.

2.3 Open Innovation

The open innovation paradigm was popularized by Henry Chesbrough (2006) and treats R&D as an open system where the company resorts more broadly to external sources of innovation or to outlicensing initiatives for its own technologies. The adaptation of this paradigm to the context of scientific research is called open science. For example, the crowdsourcing is applied in scientific domains.

The crowdsourcing refers to "*the outsourcing of tasks to a crowd that consists of a decentralized, dispersed group of individuals in a knowledge field or area of interest*" (Schildhauer and Voss, 2014). The crowdsourcing allows to collectively pool, aggregate, group and classify data. It may take different forms. Let see three examples: DBpedia.org, Observations.be and Amazon Mechanical Turk (www.mturk.com):

- Dbpedia.org is the core of the semantic Web and is based on Wikipedia, an online encyclopedia whose the content is generated by the users (Auer et al., 2007; Viseur, 2013).
- Observations.be is based on a network of benevolent passionate skilled lovers of nature to gather observations related to fauna and flora in Belgium and Netherlands.
- Amazon Mechanical Turk is a generic crowdsourcing platform that is frequently used by scientific community to access panels or processing more or less significant volumes of data (Paolacci, 2010).

The collaborative practices are also fostered by the opportunities offered by the new online tools inspired by Web 2.0, such as the academic social networks (e.g., ResearchGate.net) or the online writing tools (e.g., Google Documents).

2.4 Legal Necessity

Some funding organizations require the researchers to publish in open access and release open research data. For example, European Commission define in his H2020 program a list of domains, called Open Research Data Pilot, for which a wider access to publications and data is promoted (EC, 2014). The decision is motivated by the willingness to improve the quality of results), to foster collaboration and avoid duplication of effort, to accelerate innovation (time-to-market) and involve citizens and society.

2.5 Practical Necessity

Some research fields require the use of large amount of data. It is for example the case in medicine where numerous researches have a statistical basis. The quality and the amount of data become very

important, in particular where the number of tested factors increases. In consequence, some researchers build up networks in order to pool the data and acquire sufficient sample sets (Floca, 2014).

In other disciplines, the researchers face to data monopolies. Thus the analysis of the World Wide Web often resorts on commercial search engine due to the cost of collecting data covering the entire Web. Unfortunately the researchers face to poorly documented indexes, secret algorithms and changing terms of use. In consequence, Kilgarriff (2007) considers "googleology" as a bad science and proposes an academic-community alternative that consists in "*downloading and indexing substantial proportions of the World Wide Web, but to do so transparently, giving reliable figures, and supporting language researchers' queries*". Open data research can thus be a counterweight to data monopolies.

2.6 Notoriety

Open access can increase the number of citations (Fecher and Friesike, 2014). Yet citations and references are the most obvious forms of positive feedback in the scientific community. Moreover, a higher number of citations conducts to a greater likelihood of being quoted in the future, what Fries (2014) compares to the "Rich get Richer" phenomenon. Thus, according to Piwowar et al., (2007), sharing detailed research data is associated with increased citation rate, independently of journal impact factor, date of publication, and author country.

2.7 Ideology

Fecher and Friesike (2014) identified some schools promoting a wide access to the product of the research (e.g., publications and scientific data). The authors identified five schools (democratic, pragmatic, infrastructure, public and measurement) often promoting open data for various reasons such as the goal to make the knowledge available for everyone or the improvement of scientific efficiency through collaboration.

3 ISSUES AND SOLUTIONS

3.1 Traditional Practices

The open data research questions the scientific practices. Indeed the modern science that came to life in the 17th century is based on the professionalization and the institutionalization of the knowledge creation (Bartling and Friesike, 2014). The opening of project researches to external world is a move on traditional practices often based on individual work or dense collaborations in small scientific distributed teams. However the increasing complexity of the problems to be solved needs to join the efforts and consider multi-expert works in order to find solutions (Fecher and Friesike, 2014).

3.2 Scientific Publishers Reluctance

The scientific publication is an economic sector characterized by rising concentration. Thus four big players control about 60-70% of journal titles worldwide (Sitek and Bertelmann, 2014) and benefit on strong position. Some publishers defend their copyright aggressively and oppose new scientific practices that consist in publicly disseminating the researches with their source code, the data structures, the experimental design, the parameters, the documentation and the figures (Stodden, 2009).

Some researchers try to regain the control of editorial process and ensure competition between scientific publishers or alternatively campaign against the business practices (e.g., "The Cost of Knowledge" initiative) in order to facilitate the access to scientific knowledge. The pressure encounters success as the adoption of open access by major publishers shows (e.g., Springer Plus).

3.3 Data Property

The publication of open research data require to choose a license defining the rights and duties of the licensors and the licensees. Fortunately the Open Knowledge Foundation (okfn.org) conducts a project, called Open Data Commons (opendatacommons.org), aiming to offer set of legal tools for providing and using open data. They include three distinct licenses: the Public Domain Dedication and License, the Attribution License and the Open Database License (Miller et al., 2008; Penev et al., 2009). The first one (PDDL) is a public domain license for data and databases. The second one (ODC-By) protects the paternity and the third one (ODC-ODbL) adds a ShareAlike clause that is similar to copyleft in the free software.

The frequent use of Creative Commons for the protection of databases is sometimes criticized. The Belgian and French versions of Creative Commons, unlike the US version, contain a clause relating to

the protection of databases. As Creative Commons licenses are built on copyright, the licensed object must be a creative work (Miller et al., 2008). Neither databases nor data are creative work. In addition, the laws relating to databases differ between the European Union and the United States. The first protects databases (see "Directive 96/9/EC", europa.eu), but the latter refused to vote on a similar bill in 1991. However, the use of CC licenses for data remains recommended by Creative Commons Foundation herself (e.g., CC0).

The choice of a standard open data license is recommendable when the data are not critical one and can be globally published. The situation is more complex with sensitive data covered by specific legislations (e.g., private data or medical data). Solutions such as anonymization techniques are existing for private data but can be bypassed by reindentification techniques. The latter are facilitated by the availability of data that allow to cross information. The publication may also be hindered if the data are copyrighted by a third party or protected under confidentiality agreement.

3.4 University IPR Policies

Universities are gradually encouraged to diversify their sources of revenues by the sales of licenses (Intellectual Property Rights), the research contract signature, the creation of spin-off companies and the commercial exploitation of new inventions. Geuna and Nesta (2006) highlights consequences including a substitution effect between publishing and patenting for younger researchers and a negative impact on open science practices in the form of increased secrecy.

Current public policies in favour of open access and open research data may redress the balance between knowledge sharing and business development in universities.

3.5 Data Semantic

The pooling of data may benefit on features facilitating the crossing of data. The semantic Web concept deals with that issue. It leads to the emergence of Linked Data / Linked Open Data (LOD) concepts and the creation of several standards (Bizer, 2009; Miller et al., 2008). Thus the data related to entities are structured in RDF (Resource Description Framework). The datasets can be queried in SPARQL (SPARQL Protocol and RDF Query Language). DBpedia (dbpedia.org), the semantic version of Wikipedia, is a concrete example (Auer et al., 2007; Viseur, 2013).

However, even with semantic standards, the researcher faces the risk to compare apples and oranges. The open data research aggregation supposes to define the semantic of data that were collected in the linked databases (e.g., metadata) and the ways the data were collected (e.g, documentation of heterogeneity over time in chronological series).

3.6 Data Quality

The publication of research results with data and source code exposes the researchers to a more detailed review of their studies. The availability of quality toolkits could foster the researches to publish their data without risking their reputation.

On the other hand the researchers may benefit on the data quality to gain positive reputation, as it works in open source software. Thus the open source software is considered as a highly individualist phenomenon that is characterized by a reputation-based culture (e.g., peer recognition) (Feller and Fitzgerald, 2002). So talented researchers should benefit on the visibility offered by high quality open research data projects.

However, the researchers could benefit on setting up data quality checking tools on the same model as software quality. Beyond the specific needs of data researchers, the development of data quality standards and norms (e.g., ISO 8000) could help the structuring of those tools. Researchers may resort on existing reviews about data quality problems and tools (e.g., Barateiro and Galhardas, 2005).

3.7 Data Infrastructure

The researchers need public repositories allowing the publication of their open research data.

We could compare the context with open source software. If the open source pioneers could accept a simple Web hosting to publish an archive (e.g., .zip or .tar files) with their source code, the practices quickly evolved to dedicated integrated tools that were called forges (e.g., Sourceforge or Github). The software forges offered powerful tools in order to publish, share and concurrently access the source code for reading or updating (e.g., CVS, SVN or Git), discuss about the project and document the defects that were found in the software with bug tracker (e.g., Bugzilla).

The researchers need the same kind of tool for open research data. Some tool for general purpose has been already published (e.g., The Datatank) but their compliance with researchers requirements must

still be deepened. Indeed the open research data has its own constraints such as the usability for non expert users in case of citizen involvement or the capacity to manage large amount of data in case of big data projects.

Beyond the software issue the data infrastructure suppose the ability to fund storage, bandwidth and processing power over time. Cloud computing offers on-demand resources but may expose researchers to lock-in issues, through proprietary interfaces or data formats (Viseur et al., 2014).

4 CONCLUSIONS

In that research, we identified seven motivations for researchers to open research data and we also discussed seven issues (see Table 1).

We assume the researcher's competences should evolve to include the ability to create and animate communities around research projects based on open source and open data. The benefits are various for universities and researchers, e.g., citations, feedbacks or collaborations. Thus the university technology transfer office (TTO) could support open source and open data initiatives, as they may accelerate innovation and contribute to improve the university image. The help may encompass community issues, but also IPR (e.g., licenses). The provision of dissemination tools could be taken in charge by universities, as they would usefully complement the already existing institutional repositories where the affiliate researchers must deposit their publications, including reports, talks, conference proceedings or teaching materials.

We highlight an important issue related to the data infrastructure necessary for sharing data. In practice some generic tools exist for general open data initiatives but they are not designed for open data research. Moreover the open source software show that the collaborative tools progressively gain maturity and allow to manage the software source codes (versioning), the bugs and the documentation. The most advanced tools allow the continuous integration and automatically execute some checks on source codes, resulting in increased software quality.

Future works should be conducted in order to identify the tool chain allowing the continuous integration of open research data (data forge). The study of projects such as Wikipedia (collaborative online encyclopaedia) or OpenStreetMap (collaborative map of the world) should bring valuable findings. Future data forges should support the data acquisition (including wrappers, metadata and semantics), the data storage, the data analysis (including visualization) and the data export. They could be based on the results of various research programs, where executable workflows have already been settled up for specific purposes (e.g., DataOne; see (Reichman et al., 2011)), or the existing open source applications (e.g., Apache Taverna).

Table 1: Summary of the findings.

Motivations	
Compliance with open access practices	
Needs for reproducible research	
Adaptation of open innovation principles to science	
Legal necessity	
Practical necessity	
Search for notoriety	
Respect of an Ideology	
Issues	**Solutions**
Traditional practices in scientific community	Collaboration as a response to increasing complexity of the problems to be solved
Scientific publishers' reluctance to open access to papers and data	Regain control over editorial process, pressure on scientific publishers
Data property	Use of standard licenses for sharing contents (e.g., Creative Commons)
University IPR policies	Setting up public policies promoting open access and open research data
Data semantic	Use of semantic Web standards, use of metadata, documentation of datasets
Data quality	Automatic testing tools and their integration with "data forges"
Data infrastructure	Development of "data forges" suitable for open research data

REFERENCES

EC (European Commission). Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 (Version 1.0), 11 December 2013.

Auer, Sören, Bizer, Christian, Kobilarov, Georgi, et al. Dbpedia: *A nucleus for a web of open data*. Springer Berlin Heidelberg, 2007.

Barateiro, José, Galhardas, Helena. *A Survey of Data Quality Tools. Datenbank-Spektrum, 2005, vol. 14*, no 15-21, p. 48.

Bartling, Sönke, Friesike, Sascha. Towards another scientific revolution. In: Opening Science. Springer

International Publishing, 2014. p. 3-15.

Bizer, Christian. *The emerging web of linked data. Intelligent Systems, IEEE, 2009*, vol. 24, no 5, p. 87-92.

Claerbout, Jon F., Karrenfach, Martin. Electronic documents give reproducible research a new meaning. In: 1992 SEG *Annual Meeting. Society of Exploration Geophysicists*, 1992.

Chesbrough, Henry, Vanhaverbeke, Wim, West, Joel (ed.). *Open innovation: Researching a new paradigm*. Oxford university press, 2006.

Eaton, John W. Gnu octave and reproducible research. *Journal of Process Control, 2012, vol. 22, no 8*, p. 1433-1438.

Fecher, Benedikt, Friesike, Sascha. Open Science: one term, five schools of thought. In : *Opening Science. Springer International Publishing, 2014*. p. 17-47.

Feller, Joseph, Fitzgerald, Brian. *Understanding open source software development.* London : Addison-Wesley, 2002.

Floca, Ralf. Challenges of Open Data in Medical Research. *In : Opening Science*. Springer International Publishing, 2014. p. 297-307.

Geuna, Aldo, Nesta, Lionel JJ. University patenting and its effects on academic research: *The emerging European evidence. Research Policy, 2006, vol. 35, no 6,* p. 790-807.

Jomier, Julien, Bailly, Adrien, Le Gall, Mikael, et al. An open-source digital archiving system for medical and scientific research. *Open Repositories, 2010, vol. 7.*

Kilgarriff, Adam. *Googleology is bad science. Computational linguistics, 2007, vol. 33, no 1*, p. 147-151.

Kauppinen, Tomi, de Espindola, Giovana Mira. *Linked open science-communicating, sharing and evaluating data, methods and results for executable papers*. Procedia Computer Science, 2011, vol. 4, p. 726-731.

Lindman, Juho, Tammisto, Yulia. Open Source and Open Data: Business Perspectives from the Frontline. *In : Open Source Systems: Grounding Research.* Springer Berlin Heidelberg, 2011. p. 330-333.

Miller, Paul, Styles, Rob, Heath, Tom. *Open Data Commons, a License for Open Data. LDOW*, 2008, vol. 369.

Murray-Rust, Peter. *Open data in science. Serials Review, 2008, vol. 34, no 1, p*. 52-64.

Paolacci, Gabriele, Chandler, Jesse, & Ipeirotis, Panagiotis G. *Running experiments on amazon mechanical turk. Judgment and Decision making, 2010, vol. 5, no 5*, p. 411-419.

Penev, Lyubomir, Sharkey, Michael, Erwin, Terry, et al. *Data publication and dissemination of interactive keys under the open access model. ZooKeys, 2009, vol. 21,* p. 1-17.

Piwowar, Heather A., Day, Roger S., et Fridsma, Douglas B. *Sharing detailed research data is associated with increased citation rate. PloS one, 2007, vol. 2, no 3,* p. e308.

Reichman, O. J., Jones, Matthew B., et Schildhauer, Mark P. *Challenges and opportunities of open data in ecology. Science, 2011, vol. 331*, no 6018.ank-Spektrum, 14(15-21), 48.

Schildhauer, Thomas, Voss, Hilger. Open Innovation and Crowdsourcing in the Sciences. *In Opening Science. Springer International Publishing, 2014*. p. 255-269.

Sitek, Dagmar, Bertelmann, Roland. Open Access: A State of the Art. In*: Opening Science. Springer International Publishing, 2014.* p. 139-153.

Stodden, Victoria. The legal framework for reproducible scientific research: *Licensing and copyright. Computing in Science & Engineering, 2009, vol. 11,* no 1, p. 35-40.

Teif, Vladimir B. Science 3.0: *Corrections to the Science 2.0 paradigm.* arXiv preprint arXiv:1301.2522, 2013.

Vandewalle, Patrick et al. *Reproducible research in signal processing. Signal Processing Magazine, IEEE, 2009*, vol. 26, no 3, p. 37-47.

Viseur, Robert, Charlier, Etienne, Van de Borne Michael. Cloud Computing and Technological Lock-in: *Literature Review, Data Technologies and Applications*, Vienna, Austria, 2014.

Viseur Robert. Extraction of Biographical Data from Wikipedia, Data Technologies and Applications, Reykjavik, Iceland, 2013.