September 2016

Open Data, Challenges towards implementation, and possible solutions.

Ian Mulvany

eLife Sciences, Sage Publications

Abstract: *Calls in favour of Open Data in research are becoming overwhelming. They are at national (RCUK 2015) and international levels (Moedas 2015, Boulton et al. (2012), The Netherands EU Presidency (2016)). I will set out a working definition of Open Data and will discuss the key challenges preventing the publication of Open Data becoming standard practice. I will attempt to draw some general solutions to those challenges from field specific examples.*

Open Data is …

Open Data is Findable, Accessible, Interoperable and Reusable (M. D. Wilkinson et al. 2016). Making data available is key to support reproducibility, but by far the greatest benefit comes when data can be built upon. This truly assists with the advancement of knowledge. When data is reused there is an immediate return on the investment used for the creation of the original data.

How should we speak of the challenges for Open Data?

(Goodman et al. 2014) lay out ten rules for the care and feeding of scientific data. Were all of these rules to be adhered to by all researchers, we would have as good an Open Data ecosystem as we could wish for. Let us look at what might be preventing us from adopting these key practices. To streamline the discussion I have grouped the ten rules into three core challenges.

Core Challenge One: Competence in Working with Data.

This challenge is addressed by the following rules:

- Rule 1. Love Your Data, and Help Others Love It, Too.
- Rule 3. Conduct Science with a Particular Level of Reuse in Mind.
- Rule 4. Publish Workflow as Context.

Data that is well described and well documented and that follows standards appropriate for its discipline, is more likely to be interoperable with similar data. Such data is more useful than were it to not follow

these principles.

A number of potential issues stand in the way of generating data of this level of quality.

- Basic researcher familiarity with good data practice can be low.
- Keeping track of what the researcher did at the point of data capture can be complicated, requiring later reconstruction when tagging data.
- New scientific tools coming to market sometimes create proprietary data formats as output formats.
- Researchers sometimes create custom data formats for their research (Figshare hosts over 100 distinct mimetypes).
- Some data is heterogeneous, bringing together multi-varied data across many dimensions, in such a way that only the researcher who created that data understands how to unpick it.

Most of these issues have all been successfully addressed within specific domains or communities.

In order to improve researcher skills with working with code, Software Carpentry has reached over 120,000 students (Wilson 2016). They conduct two-day workshops instructing researchers on the basics of how to work with software. They have created a sister organisation whose aim is to do the same, but with data management - Data Carpentry. This is a ground-up effort that is being sustained by the good will of the communities within which these courses happen, and nicely demonstrates the appetite for improved skills amongst software and data intensive researchers.

Many fields have specific standards for data description, and these should be used where they exist. Most core disciplines have appropriate data repositories, but even between similar fields, a lack of harmonisation of data standards can be an issue. The ISA TOOLs (Sansone et al. 2012) initiative can help significantly with creating interoperable data standards in the life sciences, and this kind of data interoperability effort is a good example for other fields that are looking for similar interoperability.

New microscopes have frequently created new data formats. To aid with instrument interoperability the microscopy community created the OMERO (http://www.openmicroscopy.org/site/products/omero) framework, a set of standards and software tools, that supports interoperability across over 140 different image formats.

For keeping track of what happens at the point of data collection, smart tooling is an important advance. Digital lab notebooks have a place to play in this. Project Juypter (http://jupyter.org) offers a digital notebook that supports collaboration, computation, versioning and dissemination of scientific results. Tools can be configured to automatically annotate data at the point of data capture. Rinocloud (https://rinocloud.com/) offers a service that can act as a digital hub bringing together data from multiple machines, into an auto-updated lab notebook.

For heterogeneous datasets capturing metadata about the workflow can be as important capturing the data itself. In the life sciences workflow tools such as Galaxy(Afgan et al. 2016) are being increasingly used. The journal Gigascience (http://gigascience.biomedcentral.com) now hosts published Galaxy workflows on a sister site GigaGalaxy (http://gigagalaxy.net/workflow/list_published) .

Another route for creating compatible data is to make use of data standards bodies. The Open Annotation working group created a data format with a high degree of usage in the digital humanities, and ensured interoperability and openness of that data format through an open standardisation process that led the format becoming a World Wide Web Consortium standard.

Recommendations for core challenge one.

There exist many tools and resources for learning good data management practice. Communities are self-organising training to equip themselves with the techniques needed for working in data intensive research.

Skills learnt early in a career can form the basis for ongoing improvement and learning.

Recommendations for dealing with skills gaps are:

- Coordinate training programs to adopt best practice for data management skills.
- In these training programs include an introduction to tooling that can take the burden off of the researchers for managing their data.
- Encourage the publication and dissemination of good standards and workflows so that researchers can learn by example.

Core Challenge Two: Appropriate Infrastructure for Open Data.

This challenge is addressed by these rules:

- Rule 2. Share your data online with a Permanent identifier.
- Rule 5. Link Your Data to Your Publications as Often as Possible.
- Rule 6. Publish Your Code (Even the small bits).
- Rule 8. Foster and use data repositories.

Data that has an identifier, a stable home, and is well curated is data is easily findable and accessible. This requires the existence of appropriate infrastructure (Geoffrey, Jennifer, and Cameron 2015) for that particular kind of data, be that technical infrastructure (for hosting and issuing of identifiers) or organisational and human infrastructure (for curation, annotation and preservation of the data).

Some challenges that exist around the creation of good infrastructures include:

- Some fields are experiencing an explosion of data and do not have field-wide infrastructural support for their data.
- Some data is being created at a scale beyond the ability of even the best infrastructures to store that data.
- Some data does not fall neatly into a subject specific repository.
- Researchers are equally reticent to share code as they are to share data, but code sharing has some unique challenges, such as dependency management and software quality.
- Some data needs to be treated with extreme care owing to privacy concerns, and privacy infrastructure is in its infancy.
- Until recently there has been confusion around how to give credit for data contributions within the literature.

Good infrastructure does exist for data in many domains of research (e.g. high energy physics (CERN 2009), astronomy, genomics (Benson et al. 2013, Berman (2000)), however there are emerging domains for whom lack of good infrastructure is becoming a critical problem (e.g. high throughput and resolution microscopy, conectomics (Lichtman, Pfister, and Shavit 2014), computational social science). These domains share a common pattern where the tools used have reached new levels of sophistication and as a result data generation from these tools has expanded at a rate that is much faster than had been anticipated within their fields. This is leaving these fields in a momentary state of data crisis. A solution to these kinds of crises is to encourage the funding of data infrastructure, however whether this should be done on a project level, institution level, national level, or even international level remains an open and unresolved question. Building on many reports on the challenges of data infrastructure (Knowledge Exchange 2016) recommends taking into account the full research cycle when thinking about funding infrastructures, and to look to the US where the NSF has a dedicated program for the creation of shared data-centric cyber infrastructure (http://www.nsf.gov/cise/aci/cif21/CIF21Vision2012current.pdf) .

Even for data at large scale where good infrastructure does exist, it is often not possible to preserve all of the data. It is instructive to look at how high energy particle physics deals with it's data storage requirements. CERN (CERN 2009) outlines four levels of data preservation

1. Retain only the publication that the data ended up generating.
2. Preserve the data in a simplified format, this might be for outreach or training purposes.
3. Preserve the analysis software and specification of the data format.
4. Preserve the full reconstruction level data, and possibly some of the original data.

It is understood that much of the primary data coming off the detector will have to be discarded, and so their data preservation framework allows them to make decisions on what to keep based around the expected future uses of that kind of data.

It may be possible with other emerging high resolution data sources to also find ways to make decisions around whether we can preserve certain artefacts that are of lower dimensionality of the original source data. For example in conectomics one begins with high resolution images of brain slices, and a full 3-D image of a brain can be on the order of petabytes of data, however the final network diagram showing the interconnection of the neuronal scaffolding of a brain will be many orders of magnitude smaller than the original images.

Another strategy, where large scale data is concerned, is to look to use peer to peer systems for data sharing. The tool dat data (http://dat-data.com) uses a bit-torrent like protocol to allow the creation of a pool of nodes for sharing data. It has been successfully used to overcome network bottlenecks in the sharing of genomic data across Sudan.

About 70% of the lifetime cost is incurred on first write to disk, given the current rates at which the prices of long term data storage is dropping (in 1980 1GB of data storage cost $193,000, that has dropped to $0.03 in 2015 (komorowski 2015)), the question of which data sets do we need to make strategic decisions around when it comes to large storage costs will probably always be with us, but our view on the kind and size of what that data is will constantly be changing.

In terms of numbers of data sets, most researchers producing data are not producing data at large scale. The questions around how they can create good identifiers for their data, and how they can find appropriate locations to deposit their data, are equally important as for those of large data. For subject specific data there usually exists a subject specific repository. The Registry of Research Data Repositories (http://www.re3data.org) lists over 1500 research data repositories. The journal Scientific Data also maintains a more curated list (http://www.nature.com/sdata/policies/repositories) . For data that does not naturally find a home in one of these repositories there are also generic data repositories such as Figshare (https://figshare.com) , Zenodo (http://zenodo.org) , DataDryad (http://datadryad.org) , Imeji (http://imeji.org) and Github (https://github.com) . Each of these will allow a researcher to post a public version of their data with an appropriate identifier, usually a DOI. The main challenge here is in increasing awareness amongst researchers about these resources.

Data is often derived as an output of some analysis pipeline. For true reproducibility and reusability the software that was created to analyse the data also needs to be made openly available. This comes with some specific challenges around how to preserve that code, how to ensure that the code can run for other researchers, and how to manage code dependencies. The ENCODE project (Hong et al. 2016) tackled these problems by making virtual machines available that include all of the project data and software. Increasingly in the commercial software world treating all software and hardware dependencies as code is becoming standard practice, and this allows entire software stacks to be reproducibly built from a descriptive formula (K. Morris 2016).

Privacy of data is a critical issue, however the number of domains in which this is a concern is a small subset of all research domains that are producing data, so I caution that discussion around privacy do need to be taken seriously, but they should not dominate the debate around policies for making data available

where privacy is not a concern. The mantra “Open Where Possible” should be followed.

One last critical piece of infrastructure that is required for a healthy Open Data ecosystem is the ability to give credit to data. Within scholarly publishing this is done through enabling data citation. Between 2012 and 2015 a FORCE11 (https://www.force11.org) working group created the data citation principles (Altman et al. 2015). All stakeholders in the scholarly enterprise should treat data as a first class citizen, and cite it appropriately. The standard XML used in scholarly publishing has also been updated to support data citation (Mietchen et al. 2015). Journals need to encourage good data citation practice, and need to ensure that data is acknowledged correctly. For example the open access journal eLife (https://elifesciences.org) requires author list any novel data that they create in the course of writing their paper, along with giving credit to any previously published data that they used as part of their investigation.

Recommendations for Core Challenge Two:

- Advocate for smart preservation of data, taking into account the full research lifecycle.
- Increase awareness of the many data repository options that already exist, and require deposition into an appropriate repository for any given data set.
- Continually review data infrastructure needs, at national and trans-national levels.
- Cite data as a first class object.

Core Challenge Three: Creating a Supporting Culture for Openness

This is addressed the remaining rules, and can be summarised by asking how we can ensure that the correct incentives are in place to support the sharing of open data.

- Rule 7. State How You Want to Get Credit.
- Rule 9. Reward Colleagues Who Share Their Data Properly.
- Rule 10. Be a Booster for Data Science.

Irrespective of how good data management skills are, or how sophisticated data hosting infrastructure is, unless there is a willingness on the part of the researcher to make their data open, then no data sharing will happen.

For the researcher they have to put time into making their data available, and they need to feel sure that this time would not be better used in helping their careers were they doing something else with that time, such as writing grant proposals, or working on new experiments. In some cases researchers may even have a misguided fear of sharing data in the belief that by do doing they will loose out on publishing potential results that they may have otherwise been able to obtain from the data set (The International Consortium of Investigators for Fairness in Trial Data Sharing 2016).

To overcome these fears researchers motivations to share data need to outweigh their misgivings towards data sharing.

There are two strategies that can be taken towards this. The first is to make data sharing mandatory in order to receive grants or to achieve publication. The second approach is to fairly reward data and software producers for their efforts on their own merits. This is more desirable, but significantly harder to do.

Mandatory data sharing is most common on the journal, funder and discipline level. I am unaware of any national mandate. Since 2014 PLOS has required data to be made available for publication across all of

PLOS journals. This has led to a significant increase in the number of articles published that also make their underlying data available. The Wellcome Trust requests that data be made available (https://wellcome.ac.uk/funding/managing-grant/policy-data-management-and-sharing) , and the Bill and Melinda Gates Foundation require all data to be made available (http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy) . From a grassroots effort the disciple of crystallography evolved to one in which data sharing is now mandatory. Ad hoc practice became codified as required practice (H. M. Berman et al. 2008). This seems to have been an example of where an initial tight-knit community recognised the value of data sharing, and as that community grew those values of data sharing grew with it.

A greater challenge is how to change implicit behaviour. In order to do this the reward system needs to be modified to ensure that researchers get appropriate credit for the creation of data and code. There is strong evidence that open publication practices reward researchers through more citations, access to better collaborations and easier adherence to funder mandates (McKiernan et al. 2016). In particular papers that make their data available garner more citations than those that don't (H. A. Piwowar and Vision 2013). Nonetheless researchers remain fixated on articles, and in particular article in "high-impact" journals, as the primary means of validating their work. Ironically it is usually researchers that are assessing other researchers on grant award panels, or hiring committees. These bodies need to be given clear unambiguous instruction to reward researchers based on their full scientific contribution and the adoption of San Francisco Declaration on Research Assessment (http://www.ascb.org/dora/) should be mandatory for all such bodies.

To support a more nuanced way of assessing research continued investigation of altmetrics (http://altmetrics.org/manifesto/) is recommended. Tools like Depsy (http://depsy.org) , which indexes research software and gives information on it's reuse, can help to highlight to impact and contribution of this software. Finding a way to do something similar for research data is critical (J. E. Kratz et al. 2015).

Making data or software specific positions available within the academy can also create career opportunities for researchers whose contributions are critical, but currently undervalued by the current research assessment system.

Recommendations for Core Challenge Three:

- Where possible, make data sharing mandatory for the receipt of grant funding or for publication of research articles.
- Support and reward fields that have good data sharing practices through infrastructure support for data repositories.
- Educate grant award committees about best practice for research assessment.
- Create funding for explicit data and software career tracks.
- Support ways to measure and reward data reuse.

Further reading

More relevant references for this topic can be found at the Open Data Challenges (https://www.mendeley.com/groups/9199581/open-data-challenges/papers/) group on Mendeley. The bibliography and source files for this short paper can be found on the Open Data Challenges (https://github.com/IanMulvany/open-data-challenges/tree/gh-pages) repository on github.

Afgan, Enis, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, et al. 2016. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update." *Nucleic Acids Research* 44 (W1). Oxford University Press: W3–

W10. doi:10.1093/nar/gkw343 (https://doi.org/10.1093/nar/gkw343) .

Altman, Micah, Christine Borgman, Mercè Crosas, and Maryann Matone. 2015. “An introduction to the joint principles for data citation.” *Bulletin of the American Society for Information Science and Technology* 41 (3): 43–45. doi:10.1002/bult.2015.1720410313 (https://doi.org/10.1002/bult.2015.1720410313) .

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2013. “GenBank.” *Nucleic Acids Research* 41 (D1): D36–D42. doi:10.1093/nar/gks1195 (https://doi.org/10.1093/nar/gks1195) .

Berman, H. M. 2000. “The Protein Data Bank.” *Nucleic Acids Research* 28 (1): 235–42. doi:10.1093/nar/28.1.235 (https://doi.org/10.1093/nar/28.1.235) .

Berman, Helen M., Allen F. H., Kennard O., Motherwell W. D. S., Town W. G., Watson D. G., Arnold E., et al. 2008. “The Protein Data Bank: a historical perspective.” *Acta Crystallographica Section A Foundations of Crystallography* 64 (1). International Union of Crystallography: 88–95. doi:10.1107/S0108767307035623 (https://doi.org/10.1107/S0108767307035623) .

Boulton, Geoffrey, Philip Campbell, Brian Collins, Peter Elias, Wendy Hall, Laurie Graeme, Onora O’Neill, et al. 2012. *Science as an open enterprise*. June. http://royalsociety.org/uploadedFiles/Royal{_}Society{_}Content/policy/projects/sape/2012-06-20-SAOE.pdf (http://royalsociety.org/uploadedFiles/Royal{_}Society{_}Content/policy/projects/sape/2012-06-20-SAOE.pdf) .

CERN. 2009. “Data Preservation in High-Energy Physics,” no. May: 1–18.

Geoffrey, Bilder, Lin Jennifer, and Neylon Cameron. 2015. “Principles for Open Scholarly Infrastructures-v1.” *Figshare*. doi:10.6084/m9.figshare.1314859.v1 (https://doi.org/10.6084/m9.figshare.1314859.v1) .

Goodman, Alyssa, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, et al. 2014. “Ten Simple Rules for the Care and Feeding of Scientific Data.” Edited by Philip E. Bourne. *PLoS Computational Biology* 10 (4). Public Library of Science: e1003542. doi:10.1371/journal.pcbi.1003542 (https://doi.org/10.1371/journal.pcbi.1003542) .

Hong, Eurie L, Cricket A Sloan, Esther T Chan, Jean M Davidson, Venkat S Malladi, J Seth Strattan, Benjamin C Hitz, et al. 2016. “Principles of metadata organization at the ENCODE data coordination center.” *Database : The Journal of Biological Databases and Curation* 2016. doi:10.1093/database/baw001 (https://doi.org/10.1093/database/baw001) .

Knowledge Exchange. 2016. “Funding research data management and related infrastructures,” no. May. http://repository.jisc.ac.uk/6402/1/Funding{_}RDM{_}{\&}{_}Related{_}Infratsructures{_}MAY2016{_}v7.pdf (http://repository.jisc.ac.uk/6402/1/Funding{_}RDM{_}{\&}{_}Related{_}Infratsructures{_}MAY2016{_}v7.pdf) .

komorowski, Matt. 2015. “A History Of Storage Cost.” http://www.mkomo.com/cost-per-gigabyte-update (http://www.mkomo.com/cost-per-gigabyte-update) .

Kratz, John E., Carly Strasser, H. Pfeiffenberger, D. Carlson, J. E. Kratz, C. Strasser, C. Tenopir, et al. 2015. “Making data count.” *Scientific Data* 2 (August). Nature Publishing Group: 150039. doi:10.1038/sdata.2015.39 (https://doi.org/10.1038/sdata.2015.39) .

Lichtman, Jeff W, Hanspeter Pfister, and Nir Shavit. 2014. “The big data challenges of

connectomics." *Nature Neuroscience* 17 (11). NIH Public Access: 1448–54. doi:10.1038/nn.3837 (https://doi.org/10.1038/nn.3837) .

McKiernan, Erin C, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, et al. 2016. "How open science helps researchers succeed." *ELife* 5. eLife Sciences Publications Limited: 372–82. doi:10.7554/eLife.16800 (https://doi.org/10.7554/eLife.16800) .

Mietchen, Daniel, Johanna McEntyre, Jeff Beck, Chris Maloney, and Force11 Data Citation Implementation Group. 2015. "Adapting JATS to support data citation." In *Journal Article Tag Suite Conference (Jats-Con) Proceedings*. National Center for Biotechnology Information (US). http://www.ncbi.nlm.nih.gov/books/NBK280240/ (http://www.ncbi.nlm.nih.gov/books/NBK280240/) .

Moedas, Carlos. 2015. *Open Innovation, Open Science, Open to the World*. doi:10.2777/061652 (https://doi.org/10.2777/061652) .

Morris, Kief. 2016. *Infrastructure as code : managing servers in the cloud*. O'Reilly.

Piwowar, Heather A., and Todd J. Vision. 2013. "Data reuse and the open data citation advantage." *PeerJ* 1 (October). PeerJ Inc.: e175. doi:10.7717/peerj.175 (https://doi.org/10.7717/peerj.175) .

RCUK. 2015. "Concordat On Open Research Data - Version 10," no. July.

Sansone, Susanna-Assunta, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, et al. 2012. "Toward interoperable bioscience data." *Nature Genetics* 44 (2): 121–26. doi:10.1038/ng.1054 (https://doi.org/10.1038/ng.1054) .

The International Consortium of Investigators for Fairness in Trial Data Sharing. 2016. "Toward Fairness in Data Sharing." *New England Journal of Medicine* 375 (5). Massachusetts Medical Society: 405–7. doi:10.1056/NEJMp1605654 (https://doi.org/10.1056/NEJMp1605654) .

The Netherands EU Presidency. 2016. "Amsterdam Call for Action on Open Science." https://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science (https://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science) .

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3 (March). Nature Publishing Group: 160018. doi:10.1038/sdata.2016.18 (https://doi.org/10.1038/sdata.2016.18) .

Wilson, Greg. 2016. "Software Carpentry: lessons learned." *F1000Research* 3 (January). doi:10.12688/f1000research.3-62.v2 (https://doi.org/10.12688/f1000research.3-62.v2) .