Editor's Introduction

Samuel A. Moore
Panton Fellow, Open Knowledge Foundation & Dept. Digital Humanities Ph.D Programme, King's College London, London, UK

Panton Fellowships

This book is the result of a year-long Panton Fellowship with the Open Knowledge Foundation and made possible by the Computer and Communications Industry Association. This is the second year that the fellowships have taken place, so far funding five early-career researchers across Europe.

Throughout the year, fellows are expected to advocate for the adoption of open data, centred on promotion of the Panton Principles for Open Data in Science (see below). Projects have ranged from monitoring air quality in local primary schools, to

How to cite this book chapter:
Moore, S. A. Editor's Introduction. In: Moore, S. A. (ed.) *Issues in Open Research Data*. Pp. 1–9. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/ban.a

transparent and reproducible altmetrics, to the Open Science Training Initiative and now this volume on open research data.

In addition to the funding and training fellows receive, the Open Knowledge Foundation is a great network of supportive, like-minded individuals who are committed to the broad mission of increasing openness throughout academia, government and society at large. I strongly encourage anyone eligible to consider applying for a future Panton Fellowship—it has been a very rewarding year.

Panton Principles

> Science is based on building on, reusing and openly criticising the published body of scientific knowledge.
>
> (Murray-Rust et al. 2010)

In 2009, a group of scientists met at the Panton Arms pub in Cambridge, UK, to try to articulate their idea of what best practice should be for sharing scientific data. The result of this meeting was a first draft of the *Panton Principles for Open Data in Science*, which was subsequently revised and published in 2010.

The Principles are predicated upon the idea that openly sharing one's research data is wholly beneficial to the progression of science. Shared data allows research to be replicated, verified, reused and remixed. But research is competitive and there are perceived disincentives that impact on a researcher's desire or ability to share his or her data. However, the culture of data sharing, and open science more generally, appeals to the collaborative side of the researcher, asking them to consider the discipline in which they work and the progression of science over the narrowly focused desire to maintain ownership of raw data and hence maintain a

competitive edge on their colleagues. This is the backdrop against which the Panton Principles were drafted: that sharing data is simply better for science.

The original four authors came from a range of scientific disciplines and backgrounds: Peter Murray-Rust, a chemist from the University of Cambridge; Rufus Pollock, founder of the Open Knowledge Foundation; John Wilbanks, then of Creative Commons and now Sage Bionetworks; and Cameron Neylon, a biophysicist formerly of the Science and Technology Facilities Council and now Advocacy Director at the Public Library of Science. Though each author was an advocate for open science, they disagreed on the best ways to share data to the community. As Cameron recounts in his blog post, Peter above all desired a practical and simple set of rules that publishers could easily adopt to encourage data sharing. There were also disagreements on the application of a share-alike clause to ensure that the products of reused data would remain openly available, though this would be at the expense of interoperability with other forms of data sharing (Neylon 2010).

In the end, the authors decided the best solution would be to recommend that, where possible, data should be released into the public domain. They did this through the creation of four simple principles that should govern the sharing of data. In my opinion these are best read as progressive, with each principle building on the previous one, so that by the end there is a clear sense for how data should be best shared to the community. These principles read as follows:

1. **When publishing data, make an explicit and robust statement of your wishes.**
   This very general point informs the researcher that releasing data into the public domain must be done with

the necessary care such that users know the data is in the public domain. For data without such a statement, reuse rights will remain ambiguous and the public domain status of the data could potentially be revoked. As the original principles indicate: 'This statement should be precise, irrevocable, and based on an appropriate and recognized legal statement in the form of a waiver or license.' (Murray-Rust et al. 2010).

2. **Use a recognized waiver or license that is appropriate for data.**
   Building on the previous point, the statement of intent should be in the form of a licence, but one that is appropriate for data. The issue of licencing is complex and will be discussed in great detail through the chapters in this book. However, as a starting point, the authors of the Panton Principles recommend that only licences appropriate for data be used, as opposed to the Creative Commons suite of licences (except CC0), or the GNU General Public Licence or other licences intended for software.
3. **If you want your data to be effectively used and added to by others it should be open as defined by the Open Knowledge/Data Definition—in particular, non-commercial and other restrictive clauses should not be used.**
   Again building on the previous point, appropriate licences should have no needlessly restrictive clauses attached to them. For example, data should not be licensed for non-commercial use only, as this prevents the data being combined with other less restrictively licensed datasets. As the authors explain: 'these [non-commercial]

licenses make it impossible to effectively integrate and re-purpose datasets and prevent commercial activities that could be used to support data preservation' (Murray-Rust et al. 2010).

4. **Explicit dedication of data underlying published science into the public domain via PDDL or CCZero is strongly recommended and ensures compliance with both the Science Commons Protocol for Implementing Open Access Data and the Open Knowledge/Data Definition.**

   Finally, the principles arrive at the conclusion that the best licence for releasing data into the public domain is either Creative Commons Zero (CC0) or the Open Data Commons Public Domain Dedication and Licence (PDDL). These licences ensure that data can be reused for commercial purposes, without a legal obligation for attribution (though a social obligation still remains), and ensure maximum interoperability and potential for reuse, in keeping with the 'general ethos of sharing and re-use within the scientific community' (Murray-Rust et al. 2010).

   The Panton Principles are founded on the idea that science progresses faster when data can be easily shared and reused throughout the community. Of course, the principles presuppose that the data is already curated to best practice (preserved in a suitable repository, available in a non-proprietary form where possible, etc.). Their intention is simply to recommend the steps you should to take to make your data truly *open*.

   The Principles themselves have so far been endorsed by hundreds of scientists worldwide, why not add your signature today at http://pantonprinciples.org/endorse/?

The Book

This book is intended to be an introduction to some of the issues surrounding open research data in a range of academic disciplines. It primarily contains newly written opinion pieces, but also a handful of articles previously published elsewhere (with the authors' permission in each instance). Importantly, the book is meant to start a conversation around open data, rather than provide a definitive account of how and why data should be shared. The book is open access, published under the Creative Commons Attribution License (CC BY), to facilitate further debate and allow the contents to be easily and widely spread. Readers are encouraged to reuse, build upon and remix each chapter; repository managers, data curators and other communities are encouraged to detach and distribute the chapters most relevant to them to their peers and colleagues.

Within the book you will find nine chapters on diverse topics ranging from content mining to drug discovery, to the everyday use of open data in a variety of subjects. A number of issues are inextricably linked to open data, such as data citation, ethics of open data, anonymization, long-term preservation and so forth. All of these issues will be dealt with in various capacities in the ensuing chapters. Finally, the contents have been commissioned so as to strike a balance between the theoretical and the practical—some chapters offer critiques of 'open' approaches or of disciplinary approaches to open data, whilst others contain useful how-to guides for researchers who are new to open data and might not know where to begin.

The book is split broadly in two halves. The first half features pieces on general issues around open data. Peter Murray-Rust and colleagues discuss the legal issues surrounding content mining and

offer a manifesto for the 'fundamental rights' of scholars to mine content based on the phrase 'the right to read is the right to mine' (Murray-Rust et al. 2014). Next, in 'The Need to Humanize Open Science', Eric C. Kansa offers a critique of open data, and openness in general, arguing that more attention needs to be focused on the broader institutional structures that govern how research is currently conducted and less on the 'narrow technical and licensing interoperability issues' (Kansa 2014). There are then two previously published pieces by Unni Karunakara and Anthony J. Williams et al. on data sharing within the Médecins Sans Frontières organization and the importance of open data in drug discovery, respectively (Karunakara 2014; Williams et al. 2014).

The latter half of the book features chapters on disciplinary approaches to open data, offering practical advice on data sharing and exploring the subject-specific issues that surround it. Sarah Callaghan's piece offers a comprehensive look at open data in the Earth and climate sciences—barriers and drivers, carrots and sticks, and an insightful case study of one author's personal experience of open data (Callaghan 2014). Tom Pollard and Leo Anthony Celi offer a similarly insightful piece on open data in health care, looking specifically at the delicate balance between patient privacy and open data and how the need to 'do no harm' can be negotiated with the move towards data sharing (Pollard & Celi 2014).

Wouter van den Bos and colleagues then offer their perspective on data sharing in the psychological sciences, making a case for the 'need of a common data sharing policy' that responds to the needs of a discipline that has so far failed to embrace openness in any real sense (van den Bos et al. 2014). Next, Ross Mounce looks at open data in palaeontology, particularly at the complicated state of licensing within the discipline and the need for researchers to

use only licences that conform to the Open Knowledge Definition (Mounce 2014). Finally, Velichka Dimitrova describes the Open Economics Principles and the need for all economics data to be 'open by default' to facilitate reproducible research and transparency (Dimitrova 2014).

The book is not meant to be a comprehensive overview of open data and there are of course absences of subjects and viewpoints. However, I do hope the contents are informative, stimulating and, most importantly, help start a conversation around issues in open research data.

Work Cited

Callaghan, S 2014 Open data in the earth and climate sciences. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 89–106.

Dimitrova, V 2014 Open research data in economics. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 141–150.

Murray-Rust, P, Molloy, J C and Cabell, D 2014 Open content mining. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 11–30.

Kansa, E C 2014 The need to humanize open science. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 31–58.

Karunakara, U 2014 Data sharing in a humanitarian organization: the experience of Médecins Sans Frontières. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 59–76.

Mounce, R 2014 Open data and palaeontology. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 151–164.

Neylon, C 2014 The Panton Principles: finding agreement on the public domain for published scientific data. *Science in the*

Open, 22 February 2010. Available at http://cameronneylon.net/blog/the-panton-principles-finding-agreement-on-the-public-domain-for-published-scientific-data/ [Last accessed 6 August 2014].

Murray-Rust, P, Neylon, C, Pollock and R, Wilbanks, J 2010 Panton Principles, principles for open data in science. Available at http://pantonprinciples.org [Last accessed 6 August 2014].

Pollard, T, Celi, L A 2014 Open data in health care. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 129–140.

Van den Bos, W, Mirjam, J and Wulff, D 2014 Open minded psychology. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 107–127.

Williams, A J, Wilbanks, J and Ekins, S 2014 Why open drug discovery needs four simple rules for licensing data and models. In: Moore, S (ed.) *Introduction to Open Research Data.* London: Ubiquity Press, pp. 77–88.