APPENDICES

TO THE REPORT

OPEN TO ALL? CASE STUDIES OF OPENNESS IN RESEARCH

A STUDY SPONSORED BY THE RIN AND NESTA

PROJECT REFERENCE RIN/P27

ANGUS WHYTE AND GRAHAM PRYOR
UNIVERSITY OF EDINBURGH
DIGITAL CURATION CENTRE (DCC)

September 2010

CONTENTS

1. Literature Review

Working openly in science

Public scrutiny of science is becoming broader and deeper; broader as 'openness' is expected across more research fields, and deeper as more of the research process is opened up. Research funders are converging around the view articulated by the OECD[1] and others that access to publicly funded research output, including data used as evidence for its conclusions, should be better enabled. Recent reports such as *Large-scale Data Sharing in the Life Sciences*[2], the JISC-sponsored *Dealing with Data*[3] and *DCC SCARP Synthesis*[4]*,* RIN's *To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs*[5] and *Stewardship of Digital Research Data*[6] provide a useful background to the current study as they profile the data-sharing landscape, policy development and common practices in various disciplines. Complementing the recent *Open science at web-scale* JISC report[7], the current study provides an opportunity to review the evidence on the impact of open working.

Open science principles and advantages are typically framed as generally applicable across the sciences[8]. Certain research fields however have been at the forefront of 'open working', notably the life sciences, chemistry and astronomy[9]. Some studies indicate benefits to research from openness, and it is evident that open data publishing has become an accepted norm for certain kinds of data in certain fields. The debate on open science is strong on matters of principle, but our brief literature review indicates the need for ongoing assessment of how research practices are being reshaped across and within research communities i.e. what parts of their research cycle are currently open? How open are they? Why should they be open, to what and whose advantage, and on what evidence?

The growing scope of open working

Open Science signifies principles of openness and transparency that have broad and intuitive appeal. Beyond that there is ongoing debate around the scope of both the 'openness', and the aspects of 'science' it should apply to. Definitions are given in the OECD report *Principles and Guidelines for Access to Research Data from Public Funding,* which had a major influence on UK Research Councils' data access policies. The report defines openness as:

> "… access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based." (p.10)

This definition asserts that research data should be accessible to researchers across national borders. It does not explicitly state that access must be public, nor without limits on its re-use. The report does however limit the scope of 'data' to specific categories, so that its principles are not

1 OECD Principles and Guidelines for Access to Research Data from Public Funding http://www.oecd.org/dataoecd/9/61/38500813.pdf (accessed Jan 12, 2010)

2 Large-scale data sharing in the life sciences: the Joint Data Standards Study Report. http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002552

3 Lyon, (2007) Dealing with Data: Roles, Rights, Responsibilities and Relationships, http://www.jisc.ac.uk/publications/publications/dealingwithdatareportfinal.aspx

4 Key Perspectives. (2010), Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study, Digital Curation Centre, Retrieved Feb 10, from http://www.dcc.ac.uk/scarp

5 Swan, A., & Brown, S. (2008, June). To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. Technical Report, . Retrieved April 30, 2010, from http://eprints.ecs.soton.ac.uk/16742/

6 Research Information Network (2008) http://www.rin.ac.uk/data-principles,

7 Lyon, L. (2009) Open science at web-scale: Optimising participation and predictive potential. Available at: http://www.jisc.ac.uk/publications/reports/2009/opensciencerpt.aspx

8 e.g Science Commons, (2008). http://sciencecommons.org/resources/readingroom/principles-for-open-science/

9 Nature (2009) Special Issue on Data Sharing Sept 9, 2009 http://www.nature.com/news/specials/datasharing/index.html

intended to apply to "research data gathered for the purpose of commercialisation of research outcomes, or to research data that are the property of a private sector entity" (p.13). Also excluded are data restricted for individual privacy, confidentiality, or for national security reasons.

The UK Research Councils have set out further principles to guide researchers in their data policies. The BBSRC for example sets out specific requirements for Data Sharing Plans, which are peer-reviewed along with the proposals they are submitted with. Their requirements recognise the 'key principles' shown in Figure 2.

Figure 2 Key principles underlying BBRSC data sharing policy[10]

The BBSRC policy, in common with other funder's data policies, recognises that data sharing practice differs between disciplines, and is constrained by the legal and ethical requirements to protect 'human subjects' data from inappropriate disclosure. The policy also recognises that:

> "Researchers have a legitimate interest in benefiting from their own time and effort in producing the data but not in prolonged exclusive use of these data. Timescales for data sharing will be influenced by the nature of the data but it is expected that timely release would generally be no later than the release through publication of the main findings and should be in-line with established best practice in the field."[11]

Returning to the more general OECD Guidelines, these clearly encourage authors of published articles to release datasets relevant to the claims made in those articles. Concepts of *open data* typically extend beyond this to include *pre-publication* data release, which is encouraged by the BBSRC among other funders. Genomics and other life science research fields have been at the forefront of pre-publication release, and in 2009 Nature published the 'Toronto Statement'. Building on previous declarations this called for the practice to be extended to other 'data intensive' fields in biology, "when there is a community of scientists that can productively use the data quickly — beyond what the data producers could do themselves in a similar time period, and sometimes for scientific purposes outside the original goals of the project"[12].

In the OECD Guidelines 'open science' relates to a broad definition of data, considered as "factual

10 Source: Kell, D. 'Digital Data to Knowledge' presentation at IDCC'09 International Digital Curation Conference Manchester, UK, 2009

11 BBRSC Data sharing policy retrieved February 12, 2010 from: http://www.bbsrc.ac.uk/datasharing/

12 Prepublication data sharing. (2009). Nature, 461(7261), 168-170. doi:10.1038/461168a

records … used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings". However this excludes physical objects and "laboratory notebooks, preliminary analyses, and drafts of scientific papers, plans for future research" (p.14).

The term '*Open Notebook Science'*[13] (Bradley et al, 2008) on the other hand adds these to the scope of open science, with chemists and biologists leading efforts to "place the personal, or laboratory, notebook of the researcher online along with all raw and processed data, and any associated material, as this material is generated"[14] Open notebook sites adopting a blog or wiki form, such as Bradley's *UsefulChem* and the biology oriented *OpenWetWare*[15]*,* include research protocols within their scope.

As well as promoting sharing of information on best practices and 'how to' experimental procedure guides, open notebook applications such as *OpenWetWare* and the UK based *LabBlog* take the form of an electronic laboratory notebook (ELN), where the 'what happened' details of research are recorded alongside experimental data. The form of 'open notebook' (e.g. blogs versus wikis) has evolved to accommodate differences in scope and laboratory practices across different fields[16]. These differences also include the degree to which details are disclosed; the overall focus here is on keeping a full record of experiments, i.e. provenance information, with a degree of choice in how much of the data and methods are shared publicly[17]. The notion of *research objects*, or "semantically rich aggregations of resources that bring together the data, methods and people involved in (scientific) investigations"[18] extends the scope of open science even further, overlapping much of e-science along the way.

The Science Commons project *Open Science Principles*[19] (see box 1) also relates to a broad spectrum of research process and product. In 2010 an informal group of open science activists launched the *Panton Principles*[20], a set of recommendations for making scientific data open that was intended to build on previous initiatives such as the 2007 Brussels Declaration on STM Publishing[21], which committed publishers of science, technology and medicine to the principle that datasets should be submitted along with journal papers and made freely accessible where feasible.

The Panton Principles draw on the still wider *Open Knowledge Definition* [22], a work of pressure group the Open Knowledge Foundation (OKF), which takes inspiration from the Open Source Software movement and Open Access publication models. The definition claims to include any form of data or 'content', and disavows any licence or other limitations on its access and redistribution. The only conditions on re-use that accord with it are a requirement for attribution. Any limits on 'fields of endeavour' are ruled out, so they "may not restrict the work from being used in a business, or from being used for military research". The definition is intended as a benchmark for licence terms. For example the Creative Commons 'CC Zero' licence with which creators waive all copyrights would be consistent with the OKF definition, while the CC-NC licence provision for 'Non-Commercial Use'[23] would fail the 'openness' test according to this definition.

Definitions of open science, as we have seen, carry with them various assumptions about what should be shared with whom and when, along with assumptions about what can legitimately be

13 Bradley, J., Owens, K. & Williams, A., (2008). Chemistry Crowdsourcing and Open Notebook Science. Nature Precedings. Available at: http://precedings.nature.com/documents/1505/version/1 [Accessed November 30, 2009].

14 Open Notebook Science - Wikipedia. Retreived January 19, 2010 from http://en.wikipedia.org/wiki/Open_Notebook_Science

15 OpenWetWare available at: http://openwetware.org/wiki/Main_Page

16 Neylon, C. (2007) 'Science in the Open' blog post and comments retrieved October 15, 2009 from: http://blog.openwetware.org/scienceintheopen/2007/08/17/the-southampton-e-lab-blog-notebook-part-2-eln-strategy/

17 Frey, J. G. (2009). The value of the Semantic Web in the laboratory. Drug Discovery Today, 14(11-12), 552-561. doi:10.1016/j.drudis.2009.03.007

18 Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010, February 22). Research Objects: Towards Exchange and Reuse of Digital Knowledge. Conference or Workshop Item, . Retrieved April 6, 2010, from http://eprints.ecs.soton.ac.uk/18555/

19 Science Commons » Principles for open science. (n.d.). . Retrieved March 30, 2010, from http://sciencecommons.org/resources/readingroom/principles-for-open-science/

20 Panton Principles. (2010). Retrieved February 19, 2010, from http://pantonprinciples.org/

21 Brussels Declaration (2007) Retrieved February 19, 2010 from http://www.stm-assoc.org/public_affairs_brussels_declaration.php

22 Open Knowledge Definition. (n.d.). . Retrieved January 19, 2010, from http://opendefinition.org/

23 Creative Commons — Attribution-Noncommercial 2.5 Generic. (n.d.). . Retrieved January 19, 2010, from http://creativecommons.org/licenses/by-nc/2.5/

withheld, and what gains technology can bring to the equation.

We have therefore adopted a broad definition for the purposes of this report. We use 'openness' as a relative term, to describe the extent to which products of any stage of the research lifecycle are openly accessible and re-usable beyond those contracted to produce them. This entails a continuum from proprietary control to the limitless reusability envisaged by (for example) the Panton Principles. We acknowledge that for some advocates the 'open' in open science refers to the absence of legal restriction on reuse, but want to also consider social and technical aspects of accessibility and re-usability.

Science Commons Principles for Open Science

Open Access to Literature from Funded Research

By "open access" to this literature, we mean that it should be on the internet in digital form, with permission granted in advance to users to "read, download, copy, distribute, print, search, or link to the full texts of 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."

Access to Research Tools from Funded Research

By "access" to research tools, we mean that the materials necessary to replicate funded research – cell lines, model animals, DNA tools, reagents, and more, should be described in digital formats, made available under standard terms of use or contracts, with infrastructure or resources to fulfill requests to qualified scientists, and with full credit provided to the scientist who created the tools.

Data from Funded Research in the Public Domain

Research data, data sets, databases, and protocols should be in the public domain. This status ensures the ability to freely distribute, copy, re-format, and integrate data from research into new research, ensuring that as new technologies are developed that researchers can apply those technologies without legal barriers. Scientific traditions of citation, attribution, and acknowledgment should be cultivated in norms.

Invest in Open Cyberinfrastructure

Data without structure and annotation is a lost opportunity. Research data should flow into an open, public, and extensible infrastructure that supports its recombination and reconfiguration into computer models, its searchability by search engines, and its use by both scientists and the taxpaying public. This infrastructure should be treated as an essential public good.

Box 1. Science Commons Principles for Open Science

Benefits, Issues and Constraints

The discussion has already touched on advantages that Open Science proponents attribute to open working. We need to consider these further and group the claimed benefits along with issues and barriers under five sub-headings:

1. Speed and efficiency of the research cycle
2. Capabilities to identify new research questions
3. Research effectiveness and quality
4. Innovation, knowledge exchange and impact
5. Research group and career development

Speed and efficiency of the research cycle

The post-doc as agent of openness

Openness in science is credited with impacts on the speed and productivity of the research cycle in the OECD Guidelines. Yet few economic studies of research consider different forms of openness. According to David et al (2009)[24], economic studies of research productivity tend to assume all forms of publicly-funded research are "open" in that the end results are made public, by contrast with privately-funded R&D. Carayol and Matt (2006)[25] for example analyse the output of a thousand faculty members of Louis Pasteur University in terms of the impact of a range of variables on the intensity of publication. They conclude that public vs. private funding has a small impact on intensity, but a relatively minor one compared to the input of 'foreign post-docs'.

This same factor - post-doctoral research funding - plays a key part in economic modelling by Mukherjee and Stern (2009)[26]. According to their approach, the economic merit of open science is that it lowers the costs of accessing knowledge from prior generations. Maintaining that advantage requires "subsidies for specialized scientific education (e.g., postdoctoral training grants)" since these "may have a multiplier effect on maintaining Open Science", leading to the conclusion that "…the viability of science depends on maintaining an upper bound on the private financial returns that are achievable through secrecy." (p.458) We review training issues under 'career development' later below, but the implication here is that the economic benefits attributable to the openness of science depend on post-docs having access to more than the published outputs of research.

Efficiencies from secondary use

Studies identifying measurable economic benefits of open data, processes or tools in research are notably lacking however. The *potential* for efficiency benefits is considered in the recent report by Fry et al (2008)[27]. They identify three areas of increased return on investment from secondary uses of data, as follows: -

- Reduced costs of collection and duplication
- Sharing the direct and indirect costs of collection (e.g. avoiding survey fatigue and thereby improving response rates)
- New uses unforeseen at the time of collection and data mining opportunities.

Each additional time a dataset is used/re-used represents a direct financial benefit equivalent to the cost of collection. Indirect costs savings include more efficient use of scarce resources used in collecting data, including research subjects and instrumentation. The main costs of re-use to be considered against such benefits include storage costs faced by a repository, and the costs of preparing the data for curation and sharing.

The 'upstream' costs of preparation

The costs of preparing data, documenting it to recognised metadata standards, are likely to be significant according to the cost modelling studies carried out in the *Keeping Research Data Safe* project by Beagrie et al (2010)[28]. These costs will be considerably higher in some fields than others. Recent case studies point to wide variations across research communities in the availability of

---

24 David, P. A., den Besten, M., & Schroeder, R. (2009). Collaborative Research in e-Science and Open Access to Information.

25 Carayol, N., & Matt, M. (2006). Individual and collective determinants of academic scientists' productivity. Information Economics and Policy, Information Economics and Policy, 18(1), 55-72.

26 Mukherjee, A., & Stern, S. (2009). Disclosure or secrecy? The dynamics of Open Science. International Journal of Industrial Organization, 27(3), 449-462. doi:10.1016/j.ijindorg.2008.11.005

27 Fry, J., Lockyer, S., Oppenheim, C., Houghton, J., & Rasmussen, B. (n.d.). Identifying the Benefits of Curating & Sharing Research Data. Retrieved May 18, 2010, from http://www.jisc.ac.uk/publications/reports/2008/databenefitsfinalreport.aspx

28 Beagrie, N., Lavoie, B., & Woollard, M. (n.d.). Keeping research data safe (Phase 2). Retrieved May 18, 2010, from http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads

suitable standards, and in the epistemological rationale for developing them in fields that have little acceptance of 'data' as an entity that can be separated from the context of its production (Lyon et al, 2010)[29].

The related 'tacit knowledge' (Polanyi, 1958)[30] aspects of research are a limit to the potential gains from data sharing, and are long known to be a barrier to scientific communication (Mukherjee and Stern, 2009)[31]. Tacit knowledge is a core component of expertise but is by definition undocumented. It is acquired through bodily experience, e.g. riding a bike, and collective social experience e.g. negotiating traffic (Collins, 2001)[32]. While it is in principle feasible to make this explicit and document it, the point is that it is normally taken-for-granted, learned through membership of a community or presence in-situ. This limits the practicality of capturing contextual information that would enable data to be re-used by a stranger, so strategies are needed for embedding metadata capture into the everyday activity through which researchers informally exchange experiences and acquire the skills that membership of their community demands (Whyte, 2008) [33]

Communication barriers to collaboration

Other economic advantages identified with openness by Fry et al (op. cit.) are the potential for "collaboration and enhanced outcomes, better education and research training, new opportunities and uses, a more complete and transparent record of 'science', potentially more sensitive and less invasive research evaluation, and greater visibility and reward" (p.79)

Collaborative effort and open working are linked in that their economic advantages and barriers are typically cited together. Barriers to collaboration have been studied extensively in relation to Virtual Research Environments or 'collaboratories'. Bos et al's study (2007)[34] of scientific collaboratories identifies 3 main barriers to scientific research moving between "informal, one-to-one collaborations, which have long been common between scientists, and more tightly coordinated, large-scale organizational structures, which are a less natural fit".

- Transferring knowledge that requires specialist expertise and may be tacit, "Scientists can often negotiate common understandings with similar experts in extended one-to-one interactions but may have great difficulty communicating what they know to larger distributed groups."
- A culture of independence that affords freedom to pursue high risk ideas and resists "controls that many corporate employees accept as normal… Scientific collaborations must work harder than other organizations to maintain open communication channels, adopt common toolsets, and keep groups focused on common goals".
- The third barrier is the difficulty of cross-institutional work crossing formal institutional boundaries, including IPR issues, since "universities often guard their intellectual property and funding in ways that hinder multi-site collaboration".

These are relevant to issues of documentation, standardisation, and control or custodianship of data which we return to under 'quality'. They also relate to legal intellectual property rights, which we return to under the 'innovation' heading.

29 Lyon, L., Rusbridge, C., Neilson, C., & Whyte, A. (2010). Disciplinary Approaches to Sharing, Curation, Reuse and Preservation: DCC SCARP Final Report to JISC. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

30 Polanyi, M. (1958). Personal knowledge: Towards a post-critical philosophy. Routledge & Kegan Paul London.

31 Mukherjee, A., & Stern, S. (2009). Disclosure or secrecy? The dynamics of Open Science. International Journal of Industrial Organization, 27(3), 449-462. doi:10.1016/j.ijindorg.2008.11.005

32 Collins, H. M. (2001). What is tacit knowledge. The practice turn in contemporary theory, 107–119.

33 Whyte, A., Job, D., Giles, S., & Lawrie, S. (2008). Meeting Curation Challenges in a Neuroimaging Group. International Journal of Digital Curation, 3(1).

34 Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., & Olson,, G. (2007). From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories. Journal of Computer-Mediated Communication, 12(2), Article 6.

Capabilities to identify new research questions

An enhanced ability to identify research problems is commonly associated with open working. The benefits from improved data access and sharing asserted in the OECD *Guidelines* (p.10) for example are mainly in the area of enabling new research. The report points to seven main routes to this, claiming that open access:

- Reinforces open scientific inquiry
- Encourages diversity of analysis and opinion
- Promotes new research
- Makes possible the testing of new or alternative hypotheses and methods of analysis
- Supports studies on data collection methods and measurement
- Enables the exploration of topics not envisioned by the initial investigators
- Permits the creation of new data sets when data from multiple sources are combined.

Most of these benefits are associated with the first listed, i.e. with 'open scientific enquiry,' while the last on the list, combining data from multiple sources, is more strongly tied to the benefits claimed for "data driven" openness (see below).

Open scientific enquiry and Merton's norms

Recent studies, for example David et al (op cit, 2008)[35], relate the benefits of 'open scientific enquiry' to Robert Merton's 1940's work on 'the normative structure of science'. These norms are:

- Communalism: the collective pursuit of knowledge through disclosure of data and methods
- Universalism: the necessity for low barriers to entry into scientific work and discourse, to offset conformity of opinion
- Disinterestedness: the neutrality of researchers in relation to the nature of the knowledge they contribute and its impact
- Originality: as the incentive for collegiate reputation and reward, and for disclosure
- Scepticism: as the appropriate attitude towards 'priority claims', i.e. of originality, and the basis for cooperating with scrutiny of those claims by peers.

David et al describe Merton's norms as "a clearly delineated ethos to which members of the academic research community generally subscribe, even though the individual behaviours may not always conform to its strictures" (2009, p.3). That ethos also underlies the advantages identified for preserving data to be reused for new enquiry, on the principle that data is an economic 'public good', one whose value is enhanced rather than diminished by wider use. Beagrie et al, for example state that "Access to research data can catalyze further work that creates value in a variety of ways, including reinforcing or corroborating earlier inferences; expanding on the foundations laid by earlier work; and even re-purposing the data in ways that could not have been foreseen" (2008, p.17).

Departures from the open norm

Mismatches between Merton's norms and actual scientific practice have been a recurring theme in the field of science and technology studies. For example, there are limits to 'disinterestedness' as inputs to research agendas include market, policy and other social influences, with the result that the relationship of scientific inputs to outputs is 'non-linear'. Also as ongoing research involves exchanges between researchers and private and public enterprises, often through informal networks (Martin and Tang, 2007, Meagher and Lyall, 2009)[36][37], there may be many degrees of disclosure

35 David, P. A., Besten, M. D., & Schroeder, R. (2008). Will E-Science Be Open Science? SSRN eLibrary. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1317390

36 Martin, B. R., & Tang, P. (2007). The benefits from publicly funded research. Science Policy Research Unit, University of Sussex.

between private and public. Studies of day-to-day work performed in laboratories show how this is enmeshed in networks of interested actors, intermediaries and instruments that shape what research questions can legitimately be asked and how answers are validated (e.g. Latour and Woolgar, 1986)[38].

While the potential to stimulate new lines of enquiry is clear, the extent to which open data has actually achieved this has not been widely studied. Nor has it been well established how directly reusable publicly shared data are, beyond professionally curated datasets. Evidence even in the life sciences is ambivalent. On the one hand, the development of infrastructure for sharing and curating genomic and proteomic data has driven data sharing policy and led to foundational shifts in life sciences research generally, making possible entire new data-based fields such as functional genomics and systems biology (*Large Scale Data Sharing in the Life Sciences*, 2005)[39]. Fry et al 2008 (op. cit.) for example point to the example of the European Bioinformatics Institute (EBI) public databases, whose analysis led to the discovery of 'copy-number variations' in the human genome.

On the other hand, even in the genomic and proteomic fields open data sharing is limited. Piwowar and Chapman (2008)[40] indicate that while over 90% of DNA sequences are publicly released, only 30% of microarray experiment results are. According to one reported survey (Blumenthal et al 2006) 80% of life scientists report positive experiences, yet around 40% had withheld requested data in the previous 3 years. The RIN *Case studies of researchers in the life sciences (20*09) indicate that life science research fields generally have more patchy levels of data sharing than the genomics and proteomics fields.

Data driven openness

The potential of semantic web technologies to put openly accessible data to new uses does however appear strongest in "data-driven" research, which can be found in all disciplines. The rise of "data-intensive" research (Hey et al, 2009, Atkinson and de Roure, 2010)[41] envisages fundamental changes in how research question get asked of data. The benefits of openness go beyond revisiting a dataset with new hypotheses to investigate from a new perspective or with new methods. Data mining, visualisation and simulation enable research to begin with the search for patterns, and have introduced an 'inductive' aspect to the research cycle in fields not used to asking "here is the evidence, now what is the hypothesis" (Kell and Oliver, 2004)[42]. Here the benefit of public access is to remove barriers to new forms of analysis of individual and linked datasets, by publishing them using semantic web standards to render terms and relationships in machine-readable form, and enabling open access to the search algorithms that would find new patterns (Coles and Frey, 2009)[43].

Barriers to generating new patterns from data linking and exploring new questions from existing datasets are similar to those that more generally limit the impacts of open working. Coles and Frey point to lack of disciplinary standards, guidelines, and conventions for linking data and metadata, both to publications and other parts of the research process, under-developed practices in curation and preservation, licensing issues and a range of disciplinary considerations.

The pros and cons of open working have a strongly disciplinary flavour. Differences at the disciplinary and sub-disciplinary level emerge as strongly as commonalities from recent case studies from the JISC *SCARP* project (Lyon et al, 2010), and RIN Reports *Case studies in Life Sciences*, *To*

---

37 Meagher, L., & Lyall, C. (n.d.). The invisible made visible: The role of evaluation in informing processes of knowledge exchange. Working papers (p. 11). Innogen. Retrieved from http://www.genomicsnetwork.ac.uk/innogen/publications/workingpapers/title,21156,en.html

38 Latour, B., & Woolgar, S. (1986). Laboratory life: The construction of scientific facts. Princeton Univ Pr.

39 Sinnott, R., Macdonald, A., Lord, P., Ecklund, D., & Jones, A. (2005). Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models (The Joint Data Standards Study). The Biotechnology and Biological Sciences Research Council, The Department of Trade and Industry, The Joint Information Systems Committee for Support for Research, The Medical Research Council, The Natural Environment Research Council and The Wellcome Trust. Retrieved from http://www.dcs.gla.ac.uk/publications/paperdetails.cfm?id=8109

40 Piwowar, H., & Chapman, W. (2008). Prevalence and Patterns of Microarray Data Sharing. Nature Precedings. doi:10.1038/npre.2008.1701.1

41 Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

42 Kell, D., & Oliver, S. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. BioEssays, 26(1), 105, 99.

43 Coles, S., & Frey, J. (2009, May 31). The Relevance of Linking. Monograph, . Retrieved May 18, 2010, from http://ie-repository.jisc.ac.uk/419/

Share or Not to Share: Publication and Quality Assurance of Research Data Outputs[44] and *Stewardship of Digital Research Data*[45]

Research effectiveness and quality

Five main issues stand out in discussion of the impacts of openness on research quality and effectiveness: -

1. Disclosure for validation and peer review
2. Effects on publication quality
3. Documentation of the research record
4. The role of standardisation
5. Control, custodianship and ownership issues.

Disclosure for validation and peer review

The open disclosure of research results is a long-standing principle held to ensure effective validation of science, allied to peer review before or after publication. Together they form a critically important incentive mechanism, according to the Mertonian view of science described by Caroyol and Dalles (2007):

> "…it is precisely the very action of disclosing knowledge which induces the reward (reputation or credit increase), the reward system thus creates simultaneous incentives both for knowledge creation and for its early disclosure and broad dissemination within the community."

The community of peers both performs the validation and confers the recognition deserved, with anonymity in the peer review process guarding against conflicts of interest. The implications of open data for this traditional model are uncertain but potentially immense. On the one hand the 'data deluge' and the infrastructure for pre- and post-publication data deposit offer unprecedented opportunities for scrutiny. On the other hand there is no consistent approach to the peer review of datasets, as the RIN report *To Share or not to Share:Publication and Quality Assurance of Research Data Outputs* found[46]. Concerns include the additional time costs to review data, the short supply of reviewers, and the difficulties for them to understand the data. These are reinforced by Ware's (2008) study of researchers' attitudes to peer review, where 40% of reviewers and 45% of journal editors said it was unrealistic to expect peer reviewers to review authors' data.

According to some open science advocates, broader public participation entails a radical shift in the peer review process. This draws on 'user-led' or 'citizen' science as exemplified by the US-based Galaxy Zoo project, where images of galaxies captured from the Sloan Digital Sky Survey are categorised online by contributions of amateur astronomers. For example Stodden (2010)[47] says:

> "Contributions from citizen-scientists put pressure on the very definition of scientific peer and thus on the practice of peer-review, aside from the potential increase in contributions". The incentive model that peer review offers to the professional scientist is changed to a model "… closer to that of open source software. Citizen-scientists do not seem to be seeking recognition and esteem from the scientific community through their participation, but appear to enjoy discovery and making meaningful contributions to research for its own sake".

44 Swan, A. and Brown, S. (2008) http://www.rin.ac.uk/data-publication,

45 RIN (2008) http://www.rin.ac.uk/data-principles,

46 Swan, A., & Brown, S. (2008, June). To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. Technical Report, . Retrieved April 30, 2010, from http://eprints.ecs.soton.ac.uk/16742/

47 Stodden, V. (2010). Open science: policy implications for the evolving phenomenon of user-led scientific innovation. JCOM, 9(01). Retrieved from http://jcom.sissa.it/archive/09/01/Jcom0901(2010)A05

Open data effects on publication quality

The quality assurance burden is greatest for pre-publication data deposit. The *To Share* report recommends funders and research communities should develop approaches to the formal assessment of datasets. Meanwhile it points out the key role of data centres, which "apply rigorous procedures to ensure that the datasets they hold meet quality standards in relation to the structure and format of the data themselves, and of the associated metadata". However "many researchers lack the skills to meet those standards without substantial help from specialists." For even more researchers however, data centres with such specialist help are not currently available for their domain or institution.

Some studies identify a pay-off to researchers who deposit data with their publications; a correlation between this and citation frequency. A study by Piwowar et al (2007)[48] examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The study found that the 48% of trials with publicly available microarray data received 85% of the aggregate citations. To the extent that citation counts are an indicator of quality, this implies a quality bonus from open working and a strong motivator for researchers to deposit data. However the evidence is limited; causes have not been investigated, and may be discipline related.

Similar benefits have been reported for 'team science'. A large number of co-authors on a publication, an indication of collaborative research output, is also associated with a higher citation rate and here the evidence is stronger and cross-disciplinary. A study by Wuchty et al (2007)[49] for example concluded that "teams typically produce more frequently cited research than individuals do, and this advantage has been increasing over time. Teams now also produce the exceptionally high-impact research, even where that distinction was once the domain of solo authors." (p.1036). This 'team effect' may be an alternative explanation for the higher citations of papers with deposited data.

Documentation of the research record

Open Notebook Science is for some at least a partial solution to the 'tacit knowledge' issues raised earlier. The aim is a complete research record, and capturing the data and the process 'as it happens' should provide a provenance trail; a route that can later be traced back from publication to understanding of the data and context. In *Open Science at Webscale*, Lyon (2009)[50] describes the rationale using the *Usefulchem* lab wiki as the prime example:

> "ONS provides additional valuable data and procedural information where conclusions are fully supported by evidence, which may supplement established peer review mechanisms. The examination of failed experiments is particularly useful: because the raw data is available, data outliers can be identified and tagged "do not use" together with a reason.
>
> Detailed interactions with students are possible with visual evidence and comments recorded in text and histories, providing excellent learning opportunities during the scientific apprenticeship period. An audit of an experimental process can be carried out to check procedural detail: edits and deletions are recorded so a full log is available to describe the laboratory procedure".

The role of standardisation

Standardisation in any aspect of the research cycle is of course an enabler of open working, by promoting interoperability and exchange, as many recent reports have commented. Standards in documentation, data, metadata, protocols, workflows, analysis techniques, reporting and citation procedures enable re-use of resources on an open basis. The recent review of the e-Science programme by Atkins et al (2010)[51] recommends "openness as a general policy" and allies this with

48 Piwowar, H., Day, R., & Fridsma, D. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE, 2(3), e308.

49 Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. Science, 316(5827), 1036-1039. doi:10.1126/science.1136099

50 Lyon, L. (2009). Open science at web-scale: Optimising participation and predictive potential. Retrieved May 18, 2010, from http://www.jisc.ac.uk/publications/reports/2009/opensciencerpt.aspx

51 Atkins, D. et al. (2010) Building a UK Foundation for the Transformative Enhancement of Research and Innovation. Report of the International Panel for

use of international standards as much as with licensing terms. Nevertheless the take-up of standards depends on researchers' identifying enough benefits in their work being openly interoperable to motivate their efforts to comply with them. Lack of agreement within research communities on data and metadata standards is one of the biggest barriers according to Vos et al (2009)[52]. This is in large part because "curation of existing data is not an activity that researchers would normally define as part of their role" (ibid.)

Researchers' control and data custodianship

Researcher's control over data access is another major issue for open science. Fears of the potential to lose that control are a major reason for withholding data. To a large extent this equates to the creators' desire to maximise their publication opportunities before releasing data (Hedstrom, 2008)[53], and is sensitive to publishers' mandates, especially if these are enforced (Piwowar and Chapman, 2008)[54]. Where 'human subjects' data is concerned in social and health/medical research, data creators' control of the data is both a community norm and a legal obligation (Sinnot et al 2005). While much of this data can be de-identified for public release some cannot, except at relatively high cost and risk (e.g. Whyte et al 2008)[55].

Recent moves to define codes of conduct for research (RCUK, 2009, UKRIO, 2010) place obligations on data creators that are as much about maintaining control as they are about providing reasonable access to data. These obligations are shared with their institutions, but depend on researchers' compliance. The DCC SCARP case studies (Lyon et al, 2010) and the RIN's *Case studies of researchers in the life sciences* (Pryor, 2009)[56] suggest that the outcomes will depend on whether researchers identify greater "intellectual capital" from controlling access or opening it up.

Innovation, knowledge exchange and impact

Many studies have illustrated the complex and indirect links between publicly funded research, innovation and socio-economic benefits. Influential studies include Martin and Tang's 2007 framework[57] identifying seven main 'channels' through which benefits flow from research.

- Increase in the stock of useful knowledge
- Supply of skilled graduates and researchers
- Creation of new scientific instrumentation and methodologies
- Development of networks and stimulation of social interaction
- Enhancement of problem solving capacity
- Creation of new firms; and
- Provision of social knowledge.

Martin and Tang's framework considers openness as a general characteristic of science relative to private R&D, i.e. it does not link benefits to specific parts of the research cycle being open. Recent work applying the framework to open access publishing assumes that benefits are increased along more or less all of these 'channels' according to how unrestricted the access is in terms of three dimensions: cost to use, time to publication, and licence permission (Houghton and Oppenheim,

---

the 2009 Review of the UK Research Councils e-Science Programme (Forthcoming)

52 Voss, A., Asgari-Targhi, M., Procter, R., Halfpenny, P., Fragkouli, E., Anderson, S., Hughes, L., et al. (2009.). Adoption of e-Infrastructure Services: inhibitors, enablers and opportunities. Proceedings of the International Conference on e-Social Science 2009

53 Hedstrom, M. (2006). Producing Archive-Ready Datasets: Compliance, Incentives, and Motivation. IASSIST Conference 2006: Presentations. Retrieved from http://www.iassistdata.org/conferences/2006/presentations/

54 Piwowar, H. A., & Chapman, W. W. (2008). A review of journal policies for sharing research data.

55 Whyte, A., Job, D., Giles, S., & Lawrie, S. (2008). Meeting Curation Challenges in a Neuroimaging Group. International Journal of Digital Curation, 3(1). Retrieved from http://ijdc.net/index.php/ijdc/article/view/74

56 Pryor, G. (2009, December). Multi-scale Data Sharing in the Life Sciences: Some Lessons for Policy Makers. Text.Serial.Journal, . Retrieved May 20, 2010, from http://ijdc.net/index.php/ijdc/article/view/135/0

57 Martin, B. & Tang. P. (2007). The benefits from publicly funded research. University of Sussex, SPRU - Science and Technology Policy Research. Retrieved from http://ideas.repec.org/p/sru/ssewps/161.html

2010)[58].

Partners and intermediaries

Innovation resulting from either public or private R&D entails an element of involvement of the 'end-users' of the resources produced. As we noted earlier, innovation is no longer viewed as a 'linear' process whose impact on non-academic actors comes at the end of a more-or-less predictable set of steps from funding to technology development and uptake. The final shape of research innovations depends on a range of intermediaries including retailers, media and marketing companies, telecom platform operators, advertisers, distributors and consultants and their roles of "configuring, facilitating and brokering technologies, uses and relationships in uncertain and emerging markets" (Stewart and Hyysalo, 2008)

The involvement of commercial firms in scientific research is one of the main constraints on data sharing according to a study by Blumenthal (2006). It is also a contentious area as, on the one hand, commercial partners' involvement may affect researchers' neutrality, for example in the area of clinical trials meta-analyses show that involvement of pharmaceutical companies introduces bias into publication of results. On the other hand, open working affects potential income streams. Where research is meant to lead to commercialisation or involves commercial partners, openness may conflict with earning income from licensing the rights to exploit intellectual property created through the research.

IPR threats and opportunities

A common economic view of scientific impact on innovation is that cooperative open disclosure and competitive proprietary exchange of resources exist in equilibrium. However according to some observers this is threatened by the application of IPR protection mechanisms to a growing range of objects, for example the patenting of software and genetic data, and by legislation favouring the application of 'digital rights management' technologies to materials that in educational settings would otherwise be freely re-used on 'fair use' grounds (e.g. David, 2004).

The countervailing efforts of the Creative Commons project to offer model licenses that make it easier for creators to distribute research material on a 'copyleft' basis have underpinned recent moves to make openly available collections of material that have previously been limited for re-use within higher-education, such as the Jorum repository of teaching materials. The Science Commons project and Open Knowledge Foundation also have championed public domain licensing models for data, including moves to develop machine-readable licenses and 'policy languages'. The latter aim to further reduce the barriers to integrating open datasets posed by copyright and the panoply of licensing models that limit re-use. The W3C Policy Languages Interest Group[59] provides a forum for tracking developments in this area.

Research group and career development

Open working brings into sharp focus the issues of reward and recognition that recent studies of data sharing and curation have drawn attention to. The 'bottom up' growth in new genres of data publication may be changing the division of labour around data curation, coinciding with policy-makers' moves to clarify the roles and responsibilities of researchers and institutions.

Career rewards

Recent reports on sharing and curation have made recommendations on the assessment of effort spent on 'infrastructuring' – the development of datasets, tools and standards that, while openly published, have not traditionally counted as assessable research outputs for professional recognition. The need to develop career paths for 'data scientists' and other curation roles was a key recommendation of Lyon's 2007 *Dealing with Data* report to the JISC[60]. This was shortly

58 Houghton, J. & Oppenheim, C. (2010) The Economic Implications of Alternative Publishing Models. Prometheus 28:1 , pp. 41-54

59 W3C Policy Languages Interest Group (n.d.) retrieved from http://www.w3.org/Policy/pling/wiki/Main_Page (March 25, 2010)

60 Lyon, (2007) Dealing with Data: Roles, Rights, Responsibilities and Relationships, retrieved from http://www.jisc.ac.uk/publications/publications/dealingwithdatareportfinal.aspx

followed by Swan and Brown's 2008 report on *Skills, role and career structure of data scientists and curators* [61] an assessment of current practice which addressed the "blurred" situation by identifying four main roles: "data creator, data scientist, data manager and data librarian. Their report to the RIN *To Share or Not to Share: Publication and Quality Assurance of* Research Data Outputs[62] echoed this, finding that "the lack of explicit career rewards, and in particular the perceived failure of the Assessment Exercise (RAE) explicitly to recognise and reward the creating and sharing of datasets – as distinct from the publication of papers - are major disincentives." (p.8)

Citation and attribution

The need for more effective data citation and attribution mechanisms is a related issue. The same report found that "researchers who share their data also tend to receive acknowledgements (or in some cases direct citations to the datasets themselves)." However these have not so far enjoyed parity of esteem with citations to academic papers for research assessment purposes. The limited impact and academic recognition for data publication is one of the issues driving initiatives to standardise data citation methods (Brase et al 2010)[63]. Similar issues affect other contributions to 'e-infrastructure' as Atkinson and de Roure (2010) point out:

> "… if data were given the same status then the incentive structures would favour publication of data in a re-usable fashion and encourage credit through re-use and citation. A similar argument could be applied to software and workflows. In many cases in the digital world we can automatically capture and make available the flow of IPR, e.g. by analysing interaction logs and by searching for copies."

Emergence of the 'data paper'

While measures of usage offer the prospect of new impact metrics, the 'data paper' is a recent development that is bringing the dataset further into the realms of the traditional scholarly publication. According to the Creative Commons project:

> "A data paper is a publication whose primary purpose is to expose and describe data, as opposed to analyze and draw conclusions from it. The data paper enables a division of labor in which those possessing the resources and skills can perform the experiments and observations needed to collect potentially interesting data sets, so that many parties, each with a unique background and ability to analyze the data, and may make use of it as they see fit." (Rees, 2010)[64]

Rees identifies this development with only a few fields, notable examples being in earth sciences, ecology and robotics. He also points to similarities with the concept of the 'overlay journal' in meteorological sciences, an approach to annotating a dataset with a document setting out the peer review process followed and its results (Callaghan et al, 2009).

Researcher motivations

The apparent lack of career structure begs the question of why researchers are motivated to devote effort to open data publication. Given the issues around peer review of data mentioned earlier, the disincentives appear to outweigh incentives when, as Lyon's *Open Science at Webscale* report (2009) highlights, "vulnerability to data predators and 'scooping' is a particular problem where the time lag between discovery and publication is relatively long" (p.21).

Economic studies of open source software development suggest that, if the parallels with open data publishing hold true, motivation may come from 'signalling of ability' according to Lerner and Tirole

---

61 Swan, A., & Brown, S. (2008). Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs. Retrieved from http://www.jisc.ac.uk/publications/publications/dataskillscareersfinalreport.aspx

62 Swan, A. and Brown, S. (2008) http://www.rin.ac.uk/data-publication

63 Brase, J., Farquhar, A., Gastl, A., Gruttemeier, H., Heijne, M., Heller, A., Piguet, A., et al. (2009). Approach for a joint global registration agency for research data. Information Services and Use, 29(1), 13-27. doi:10.3233/ISU-2009-0595

64 Rees, J. (2010). Recommendations for independent scholarly publication of data sets. Retrieved from http://neurocommons.org/report/data-publication.pdf

(2002)[65] studies of large-scale OSS endeavours. Signalling refers to the benefits that open source programmers receive from peer recognition (through take-up of the software) and from making their programming skills visible to an audience of future employers or investors. This promise of higher visibility to the researcher, host institution or funder is one of the benefits of data sharing suggested in Fry et al's report to the JISC *Identifying benefits arising from the curation and open sharing of research data* (2008). Others (e.g. Bitzer 2004)[66] point to the small scale and low visibility of most OSS development and identify 'intrinsic interest' as the key motivator among "highly qualified, young and motivated individuals". This refers to some combination of three factors:

- User programmers actually need a particular software solution
- The fun of play, or mastering the challenge of a given software problem, and
- The desire of belonging to the 'gift society' of active OSS programmers.

There are interesting parallels to explore with open research data, notebooks and software.

A framework for characterising open working

The *Open to All?* report describes a need to support research groups, institutions and research communities to work with a level of openness that provides advantages to them. To help identify this in the shape of data management plans and policies, a useful first step may be to characterise current research practices and assets. The framework developed for analysing examples in the study may be useful for this purpose.

Like Fry et al (2009) we find it helpful to consider two dimensions of open working:

1. The stage in the research process that sharing occurs, from the raw material at one end to published articles and datasets at the other.
2. The level of 'aggregation' of the actors involved, from the researcher and research group at one end to the public at large at the other end, with the policies and practices of funders and institutes operating between these levels.

The case studies in section 2 are organised along the first dimension, the research cycle presented in the main report. The outputs of each step in the research cycle are characterised in Table 1.

Research cycle stage	**Outputs**
Conceptualising and networking	Messages, posts, user profiles, bibliographies, resumes
Proposal writing and design	Proposal drafts, data management plans, regulatory compliance documentation, study protocols
Collecting and analysing	Raw and derived data, metadata, presentations, podcasts, posters, workshop papers
Documenting and sharing	Lab notes, research memos, study-level metadata, readme files, FAQ's supplementary information
Publishing and reporting	Conference papers, journal articles, technical reports
Engaging and translating	General articles, web pages, briefings, public exhibits, presentations

65 Lerner, J., & Tirole, J. (2002). Some Simple Economics of Open Source. Journal of Industrial Economics, 50(2), 197-234. doi:10.1111/1467-6451.00174

66 Bitzer, J., Schrettl, W., & Schr{ö}der, P. (2007). Intrinsic motivation in open source software development. Journal of Comparative Economics. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0147596706000643

Infrastructuring	Software tools, databases, repositories, web services, schemas and standards

Table 1. Research cycle stages and material outputs

The second dimension we see in terms of the six main 'degrees of openness' characterised in Figure 1. The framework shows a continuum of research materials disclosure by creators to other actors in their production and potential re-use. It begins with actors formally (contractually) creating outputs and extends to those not previously linked to them, including the general public.

We characterise the six degrees of openness as follows: -

1. Private management: sharing within a research group, where resources are organised to facilitate access and re-use by researchers within the group to at least some data or metadata on all research activity.
2. Collaborative sharing: sharing between members of a consortium established to deliver a project or programme, so that researchers employed may access data or metadata on the shared activity, e.g. on an intranet, and re-use it for their common contractual purpose.
3. Peer exchange: sharing between members of a researchers' network of peers, on the understanding that disclosure or re-use have conditions attached e.g. using a social networking web platform.
4. Transparent governance: disclosure to an external party under due process of a publicly accountable code enforced by institutions, funding bodies or government; e.g for research examination, assessment, ethical scrutiny, health and safety inspection; or where sharing is facilitated by a third party such as an archive with an institutional or funding body mandate.

Private management	**Collaborative sharing**	**Peer exchange**	**Transparent governance**	**Community sharing**	**Public distribution**
Access and use within project of single organisation. Custodian controlled Defined by employment Secure access	Access and use within consortium Multiple custodians Defined by employment Secure access	Access and use by acquaintance Defined by reputation, trust Informally operated, according to norms e.g. reciprocity	Accessible for oversight Third party verification of process Defined by voluntary code or law Standard Operating Procedures Open inspection	Access and use by community Defined by membership, voluntary registration Access or licence controls may apply	Access and use public Time limits, embargoes may apply

Figure 1 Six degrees of openness in research materials disclosure

5. Community sharing: sharing with access or re-use limited to identified and authenticated members of a research community or communities, where members are defined by their affiliation to an institution, research network and/or association or professional body. This includes for example many Virtual Research Environments, and archival resources licensed for educational access/use.
6. Public sharing: sharing where resources are made available for access by any member of the

public, at least some data or metadata on the research activity is designed to be understood by a lay audience and re-used by a designated research community, and with few restrictions - such as a limited embargo period.

The framework is not meant to suggest that degrees of openness are mutually exclusive, nor is it a prescriptive model setting out a series of stages that should be followed. For example members of a project consortium might collaboratively curate a dataset without first sharing it informally; and community sharing need not involve third party verification. Also each of the 'degrees' could itself be seen as a continuum. With its simplifications, the framework nevertheless helps to summarise examples in the evidence collected that demonstrate the varying shades of openness we found. These are summarised at the end of the report, and related to the research cycle outlined earlier.

Table 2 relates the case studies to both dimensions – research cycle stage, and degree of disclosure of relevant outputs. The same 2 x 2 matrix can be used, as in Table 3 at the end of the report, to classify the examples discussed under 'Issues and Evidence'. A similar matrix might be used to plot current practices in a research group or broader collaboration as an aid to planning where changes are desirable, and for considering the policy and technology support that would be required to accomplish them. This would mean working with the groups involved to identify the value proposition, risks, incentives and shifts in responsibility entailed in disclosing research assets a step further.

	Infrastructuring - tools services and standards	Conceptualising and networking	Proposal writing and design	Conducting and presenting	Documenting and sharing	Publishing and reporting	Engaging and translating
Public Distribution	A B C L N			A B E L	A C L	A B C	A B L
Community sharing	A C	C N	C	N			
Transparent governance	A C		C N		C	C	
Collaborative sharing	A C			C	A C		A C
Peer exchange	A C			N			A
Private management	A C			N			A

Key: A) Astronomy B) Bioinformatics C) Chemistry N) Neuroimaging
L) Language Technology E) Epidemiology

Table 2. Degrees of sharing across the research cycle in six case studies

Examples Table

Table 3 below lists examples, some of which are mentioned in the *Open to All*? case studies. The colour codes below identify the case study each relates to. Some examples appear more than once as they relate to multiple research cycle steps or levels of openness.

Astronomy | Bioinformatics | Chemistry | Neuroimaging | Language Tech. | Epidemiology

	Infrastructuring - tools services and standards	Conceptualising and networking	Proposal writing and design	Conducting and presenting	Documenting and sharing	Publishing and reporting	Engaging and translating
Public Distribution	1 4 6 7 8 10 11 13 16 17			5 11 12 13 15	1 2 4 8 19	1 20 21 22 23	13 30 27
Community sharing	4 8 27	3 9	9	14			
Transparent governance	4 8		3 8 18		19	24	
Collaborative sharing	4 8 27			8	8 4		29 25 27
Peer exchange	4 8 27						27
Private management	1 4 8 27		3				26

Table 3 Examples of sharing across the research cycle to varying degrees of openness

1 e-Crystals open access database and publication repository, enabling derivative science in crystallography: Welcome to eCrystals - University of Southampton. (n.d.). . Retrieved May 7, 2010, from http://ecrystals.chem.soton.ac.uk/

2 Datasets on human dialogue in meetings extensively reused in language technology: AMI Meeting Corpus. (n.d.). Retrieved March 7, 2010, from http://corpus.amiproject.org/

3 Pooled approach to study recruitment in neuroimaging supported by the UK Clinical Research Network: NIHR CRN CC - (n.d.). . Retrieved May 21, 2010, from http://www.ukcrn.org.uk/index/networks/uk_wide.html

4 Sky survey curated datasets available for community and then public reuse: WFCAM Science Archive. (n.d.). . Retrieved May 21, 2010, from http://surveys.roe.ac.uk/wsa/

5 AMI Meeting Corpus used in competitive evaluation of speech recognition technologies: Hain, T., El Hannani, A., Wrigley, S. N., & Wan, V. (2008). Automatic speech recognition for scientific purposes-webasr. In Proceedings of the international conference on spoken language processing (Interspeech 2008).

6 Infrastructure for open data linking in functional genomics: Zhao, J., Miles, A., Klyne, G., & Shotton, D. (2009). OpenFlyData: The Way to Go for Biological Data Integration. In Data Integration in the Life Sciences (pp. 54, 47). Retrieved from http://dx.doi.org/10.1007/978-3-642-02879-3_5

7 Effective collaboration in astronomy enabled by open infrastructure collated by European Virtual Observatory: VO-enabled scientific papers. (n.d.). Retrieved December 18, 2009, from http://www.euro-vo.org/pub/fc/papers.html

8 Effective collaboration and regulatory compliance using chemistry laboratory blogs: Frey, J. G. (2009). The value of the Semantic Web in the laboratory. Drug Discovery Today, 14(11-12), 552-561. doi:10.1016/j.drudis.2009.03.007

9 Cameron Neylon's open request for assistance on Friendfeed http://friendfeed.com/cameronneylon/9875b15c/request-for-assistance

10 Open source problem solving efficiency in astronomy: Tedds, J. A. (2009, July). Science with VO tools: the AstroGrid VO Desktop. Retrieved May 21, 2010, from http://adsabs.harvard.edu/abs/2009mavo.proc...73T

11 Shuffl: Supporting curation of small-scale research data for web publication, by gathering metadata as a product of data analysis capabilities. Retrieved May 10, 2010, from http://www.jisc.ac.uk/whatwedo/programmes/inf11/jiscri/shuffl.aspx

12 Open infrastructure in astronomy enabling multi-wavelength analysis: Nakos, T., Willis, J. P., Andreon, S., Surdej, J., Riaud, P., Hatziminaoglou, E., Garcet, O., et al. (2009). A multi-wavelength survey of AGN in the XMM-LSS field. I. Quasar selection via the KX technique. Astronomy and Astrophysics, 494, 579-589.

13 LODD – integrating data sources on Western and Traditional Chinese Medicine: Jentzsch, A., Zhao, J., Hassanzadeh, O., Cheung, K. H., Samwald, M., & Andersson, B. (2010). Linking Open Drug Data.

14 Pooling schizophrenia datasets to identify genetic causes: Stefansson, H., Ophoff, R. A., Steinberg, S., Andrea

15 Visualising geospatial and public health data for novel approaches in epidemiology: Cecchi, G., Paone, M., Franco, J. R., Fèvre, E. M., Diarra, A., Ruiz, J. A., Mattioli, R. C., et al. (2009). Towards the Atlas of Human African Trypanosomiasis. International Journal of Health Geographics, 8(1), 15. doi:10.1186/1476-072X-8-15

16 Building natural language generation tools from open source components: OpenCCG (n.d.). Retrieved March 16, 2010, from http://openccg.sourceforge.net/

17 Open source toolkit a de facto standard for statistical modelling in neuroimaging:

SPM software - Statistical Parametric Mapping. (n.d.). Retrieved May 11, 2010, from http://www.fil.ion.ucl.ac.uk/spm/software/

18 Web-based Integrated Research Application System demonstrates transparent governance of confidential data management: IRAS (n.d.). Retrieved May 21, 2010, from https://www.myresearchproject.org.uk/signin.aspx

19 Third party timestamps used to provide a transparent record of experimental provenance: Science in the open » Blogs vs Wikis and third party timestamps. (n.d.). . Retrieved December 3, 2009, from http://blog.openwetware.org/scienceintheopen/2007/08/23/blogs-vs-wikis-and-third-party-timestamps/

20 Crystallography journals enabling data publication (IUCr) Structure Reports. (n.d.). . Retrieved May 21, 2010, from http://journals.iucr.org/e/

21 VizieR Service- publishing astronomy catalogue data: VizieR. (n.d.). Retrieved January 31, 2010, from http://vizier.u-strasbg.fr/viz-bin/VizieR

22 Semantic data publication in bioinformatics: Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Comput Biol, 5(4), e1000361. doi:10.1371/journal.pcbi.1000361

23 Semantically-enhanced chemistry journal searching in: Project Prospect. (n.d.). . Retrieved Jan 11, 2010, from http://www.rsc.org/publishing/journals/projectprospect/index.asp

24 Enabling data citation and attribution, e.g. in chemistry: DataCite International Data Citation: www.datacite.org. (n.d.). Retrieved April 10, 2010, from http://www.datacite.org/

25 Astronomy tool translated for medical diagnosis: Brenton, J. ., Caldas, C., Irwin, M. ., Akram, A., Gonzalez-Solares, E., Lewis, J. ., MacCullum, P., et al. (n.d.). PathGrid: The Transfer of Astronomical Image Algorithms to the Analysis of Medical Microscopy Data. Astronomical Data Analysis Software and Systems XVIII ASP Conference Series, 411.

26 Astronomy tool commercialised for medical image analysis: MOPED: blackfordanalysis.com. (n.d.). Retrieved March 17, 2010, from http://blackfordanalysis.com/

27 AGAST Project - enabling multi-level access to datasets, from astronomy to health informatics:

Sinnott, R. O., Doherty, T., Gray, N., & Lusted, J. (2009). Semantic security: specification and enforcement of semantic policies for security-driven collaborations. Studies in Health Technology and Informatics, 147, 201-211.

28 Applying bioinformatic data-linking applications in humanities research: CLAROS Classical Art Research Online Services. (n.d.). Retrieved February 8, 2010, from http://www.clarosnet.org/about/default.htm

29 Neuroscience application of the chemistry LabBlog: NeuroHub. (n.d.). . Retrieved December 7, 2009, from http://neurohub.ecs.soton.ac.uk/index.php/Main_Page

30 Language technology tool applied across disciplines: NITE XML Toolkit - Edinburgh Home Page. (n.d.). . Retrieved May 21, 2010, from http://groups.inf.ed.ac.uk/nxt/

Interview Schedule & Topics

The interviews carried out for the case studies in the *Open to All?* report were semi-structured, i.e. the questions that follow below were used as a template, but the attention to each topic varied according to their relevance.

The schedule refers to 'evidence maps'. The methodology had been planned so that an evidence map would be produced to summarise each interview, by identifying positions taken on the main interview topics, the arguments used to support these positions and relevant examples or other evidence. An initial evidence map was drafted to refer to in the interviews. However as most participants were not familiar with the Science Commons *Principles for Open Science*, these took priority as a reference point. The approach was changed and, as the main report indicates, text summarises of the positions, arguments etc were used instead.

Interview schedule

1. Introduce study aims, talk through information sheet
2. Consent terms - permission to record / take notes
3. Roles and responsibilities for making data/tools available
4. Ref. Science Commons Principles, which aspects of research process meet these definitions,
5. Explore their views on the study questions, why they hold them and how these relate to arguments expressed in literature (referring to 'draft evidence map')
6. Ask for examples and any demonstrable evidence of open science benefits realised or sought, and barriers to this, in terms of improved research process, impacts on application/ innovation and career progression.
7. What can be done? Policy measures, how funders may enable more use of open methods in ways consistent with benefits & barriers experienced in their research domain, and to facilitate collaboration with other domains
8. Their questions or comments
9. What happens next – will draft 'evidence map' and a 1-2 page sketch of the research group based on the interviews- for checking and comment.
10. Thanks

Interview questions/prompts

1. Who is involved in sharing any data or tools online beyond the group?
2. How is that organised e.g. do you work out the approach, skills and resources required on a project-by-project basis or on some other basis?
3. How are the (ref) *Science Commons 'Principles for Open Science'* relevant to work in your field and especially to your own work?
4. Thinking of the range of datasets, tools and resources you have been involved in using or producing, are there any that go most or all the way to meeting these principles?
5. *We hope in this project to gather evidence, which might be examples from your own work or from your field that back up your views on benefits and constraints on open working.*

   Some of the questions about open working we are most interested in are summarised in this diagram or 'evidence map', which also shows a range of statements about them or positions that have been put forward.

   What I would like to do is explore your own views on whichever of these positions you want to talk about, and your reasons including any examples you can think of.

What aspects of your work would you term 'open' e.g. data you share or use, methods, infrastructure?

6. What have been the main factors in deciding how and when to open up your work so others may access it and possibly re-use it?
    - What are the differences between resources you share with colleagues you know and resources you make available more publicly?
7. Does the idea of 'data driven' science have any relevance to your work? If so, how have methods of identifying patterns or correlations from data proved useful in your work?
    - Are there examples of it leading to research questions that would not have been identified otherwise?
8. Are there tangible benefits from open working in terms of working more efficiently that you can point to from your recent projects or experience?
    - How do you weigh up the benefits against costs? E.g. availability of institutional infrastructure, funding to develop that infrastructure, available skills?
    - What research costs, if any, do you see being saved through more open working?
9. Thinking of how open working might affect research quality in your field:
    - Has your research benefited from the link between sharing data more openly and the development of standards to support sharing – has that so far been a virtuous or viscous cycle?
    - To what extent is it useful document and share research notes that describe how and when you have gathered your data or research materials? Are you familiar with the idea of 'Open Notebooks' and if so is that approach feasible or desirable?
    - Are there research questions that can be addressed more effectively though open working? Are these 'normal' research questions, or if not how do they compare?
    - (where research involves human subjects) can useful research be done with the data if it is fully anonymised?
    - Are any changes needed in how research funders or institutions govern access to 'human subjects' datasets, that would in your view help you do research more effectively? What would be the most important changes relating to openness and access?
10. Considering the effects of open working on the take-up of your research by users, or in other fields-
    - How are open working and collaboration linked in your view ? e.g. in what circumstances would open working help or hinder collaboration?

      [e.g. by signalling to others what you are capable of, or undermining others interest in working with you to use your resources]
    - How do copyright or other IPR factors affect what you can make available?
    - Has any of your research been commercialised or taken up for some non-commercial application? If so, how does having a commercialisable resource affect making the research accessible and reusable?
    - Does ethics governance count towards making the research process more open (or 'partially open')? Why?

      (e.g. regulatory scrutiny – to make the process transparent for compliance purposes- but available only to researchers through an application process?)
    - Does open science have any connections, in your views, with the idea of involving people who are interested in or affected by your research, either as specialists in something else or 'ordinary members of the public',

      for example in gathering data ('citizen science')?

or in assessing its impact ('public engagement')?

11. Considering the effects of open working on developing the research group and individual careers:
    - o If researchers early in their doctoral training need access to data and tools used by more established researchers –from where do they get access - from within the group or wider? Informally or formally?
    - o Who benefits from that, and what constraints are there? E.g. are there teaching and learning benefits from sharing data – or tools? What about beyond the group? If so how tangible are they? Any examples?
    - o Are there career rewards in terms of esteem or recognition from publishing data, tools or resources in your field? If so how tangible are these? Any examples?

      E.g. is there peer pressure/ recognition for depositing in or publishing databases, developing ontologies, standards, wiki's, blogs? Are there methodology journals that you publish in and which you can ask anyone to cite if they are reusing the data?
    - o Are there arguments/evidence for limiting access by embargoing data or resources for a period?
    - o How could funders, institutions or publishers best ensure that any data or methods you publish are attributed to you by anyone who re-uses them?
12. Where open working has potential that is not being met in your view, what should be done to address the barriers and who should do this? (funders, institutions, data archives, publishers …)?
    - o Research training?
    - o Quality assessment (HEFCE/SHEFC)?
    - o Infrastructure?
    - o Policy statements?

Information Sheet and Consent Form

What are the study aims?

The Research Information Network (RIN) and the National Endowment for Science, Technology and the Arts (NESTA) have funded DCC to carry out a number of case studies in "Open Science". These will examine what motivates researchers to work (or want to work) in an open manner with regard to their data, results and protocols, and whether advantages are delivered by working in this way. The project will investigate the perceived benefits to researchers, research institutions and funders, and the degree to which there is evidence to support these perceptions. We will also examine the disincentives and barriers to such 'open science' methods.

What do you mean by Open Science?

The pressure group Science Commons defines 'open science' in terms of principles (see box), which serve as a useful reference point, although this should not be taken as an endorsement of them by DCC or the project funders. We aim to explore what 'degree of openness' you practice or find acceptable, why, and with what consequences for research. The case studies are mainly concerned with tools, data and infrastructure. While these may underpin openly accessible literature, we are not focusing on repositories or policies for open access publication.

Science Commons Principles for Open Science

Open Access to Literature from Funded Research

By "open access" to this literature, we mean that it should be on the internet in digital form, with permission granted in advance to users to "read, download, copy, distribute, print, search, or link to the full texts of 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."

Access to Research Tools from Funded Research

By "access" to research tools, we mean that the materials necessary to replicate funded research – cell lines, model animals, DNA tools, reagents, and more, should be described in digital formats, made available under standard terms of use or contracts, with infrastructure or resources to fulfill requests to qualified scientists, and with full credit provided to the scientist who created the tools.

Data from Funded Research in the Public Domain

Research data, data sets, databases, and protocols should be in the public domain. This status ensures the ability to freely distribute, copy, re-format, and integrate data from research into new research, ensuring that as new technologies are developed that researchers can apply those technologies without legal barriers. Scientific traditions of citation, attribution, and acknowledgment should be cultivated in norms.

Who is responsible for the study?

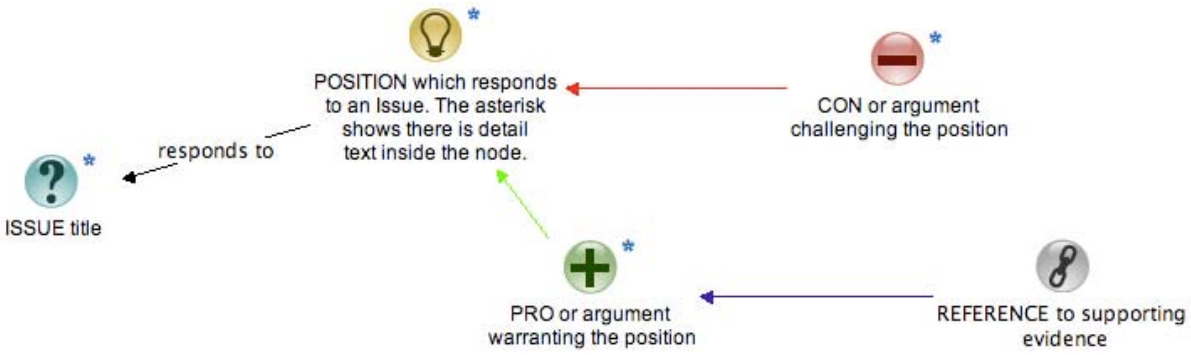
Dr Angus Whyte is carrying out the study. He is Research Officer at the Digital Curation Centre (DCC), University of Edinburgh. DCC is funded by JISC to provide a national focus for research and development into data curation issues and to promote expertise and good practice in the management of all research output in digital format. The Principal Investigator is Graham Pryor, DCC Deputy Director (contact details are at the foot of p.3).

What benefits may there be to taking part?

The case studies will reference aspects of your work that demonstrate benefits or constraints on working openly (whether wholly or partially). Provided that your Principal Investigator consents to your group being identified, taking part should bring your research activities to the attention of a wide audience, as the case studies will be publicized on the DCC, RIN and NESTA websites. You will also be contributing to the evidence base for policy makers and funders.

What information will be gathered and why?

The case studies will use interviews to gather views and experiences relating to the project aims. We would record and transcribe these and, if you agree, share these for research or educational purposes. We will also build 'evidence maps' – graphically summarizing views given by each participating research group and reasons given for holding them. The figure below shows the basic structure of these. The evidence maps and other project outputs will name research groups by field/domain and institution but need not refer to named individuals. Quotes and 'evidence maps' may be used in the final report to RIN/NESTA, in academic publications, meetings, or in presentations to non-academic audiences.



Evidence map structure

What risks and safeguards are there?

To the extent that you are identifiable, people may form opinions of you or your research group based on the representation of your views in the case study outputs. There are a number of safeguards to ensure fair representation; you will have the opportunity to view and comment on drafts, and to ensure accurate reporting of your views we hope to record and transcribe interviews. You will be asked for permission to record on any and each occasion you are involved. You are entitled to choose not to be recorded, and recording will be stopped at any point if you ask.

You can also choose not to participate further in the study at any time and do not have to give a reason for this. Transcripts and reports will refer to you by name, or if you prefer by pseudonym. Your academic role and research group will be identified with permission of you and your Principal Investigator.

Interviews or other data identifying yourself or other people by name, and not already public, will be treated as personal under the 1998 Data Protection Act and stored securely. Please note though that if your research group and institution are identified it is likely to be impossible to completely conceal your identity.

Your consent to take part in the Open Science Case Studies

You will be asked to complete the section below only when you are interviewed

---

I have read and understood the project information sheet dated 5/11/2009 ☐

I have been given the opportunity to ask questions about the project. ☐

I agree to take part in the project. Taking part in the project will include being interviewed and audio recorded. I will also be invited to comment on outputs from the project ☐

I understand that my taking part is voluntary; I can withdraw from the study at any time and I will not be asked any questions about why I no longer want to take part. ☐

Please select only one of the following two options:

I would like my name used where what I have said or written as part of this study will be used in reports, publications and other research outputs so that anything I have contributed to this project can be recognised. ☐

I do not want my name used in this project. ☐

I understand that my words may be quoted in publications, reports, web pages, and other research outputs but my name will not be used unless I requested it above. ☐

I agree for the interview data I provide to be archived at Digital Curation Centre. ☐

I understand that others may access this data only if they agree to preserve the confidentiality of that data and if they agree to the terms I have specified in this form. ☐

I agree to assign the copyright I hold in any interviews I give to this project to the University of Edinburgh. ☐

_____

Name of Participant

_____

Name of Researcher

Contact address: Researcher: Dr Angus Whyte a.whyte@ed.ac.uk Principal Investigator: Graham Pryor g.pryor@ed.ac.uk Digital Curation Centre, University of Edinburgh Appleton Tower, Crichton Street, Edinburgh EH8 9LE.