The International Journal of Digital Curation

Issue 1, Volume 6 | 2011

Open Science in Practice: Researcher Perspectives and Participation

Angus Whyte,

Curation Research Officer,

Digital Curation Centre

Graham Pryor,

Associate Director,

Digital Curation Centre

Abstract

We report on an exploratory study consisting of brief case studies in selected disciplines, examining 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. We review the policy background to open science, and literature on the benefits attributed to open data, considering how these relate to curation and to questions of who participates in science. The case studies investigate the perceived benefits to researchers, research institutions and funding bodies of utilising open scientific methods, the disincentives and barriers, and the degree to which there is evidence to support these perceptions. Six case study groups were selected in astronomy, bioinformatics, chemistry, epidemiology, language technology and neuroimaging. The studies identify relevant examples and issues through qualitative analysis of interview transcripts. We provide a typology of degrees of open working across the research lifecycle, and conclude that better support for open working, through guidelines to assist research groups in identifying the value and costs of working more openly, and further research to assess the risks, incentives and shifts in responsibility entailed by opening up the research process are needed.[1]

[1] This paper is based on the paper given by the authors at the 6th International Digital Curation Conference, December 2010; received December 2010, published March 2011.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

We report on an exploratory study conducted between November 2009 and April 2010 for UK organisations Research Information Network (RIN) and National Endowment for Science, Technology and the Arts (NESTA). This consisted of brief case studies in selected disciplines, examining 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. We review the policy background to open science, and literature on the benefits attributed to open data, considering how these relate to curation and to questions of who participates in research.

The case studies investigated the perceived benefits of utilising open scientific methods, the disincentives and barriers, and the degree to which there is evidence to support these perceptions. Six case study groups were selected to give a range of viewpoints across disciplines on the principles and practicality of working openly, as well as a range of experience in doing so. The groups were in astronomy, bioinformatics, chemistry, epidemiology, language technology and neuroimaging. The studies, involved interviews with 18 researchers from six institutions. Qualitative analysis of transcripts was supplemented by desk research and literature review.

Policy and Research on Openness in Science

Our review indicates that open science principles and advantages are typically framed in terms of science as a whole, but certain fields have been at the forefront of ‘open working’, notably the life sciences, chemistry and astronomy (Nature, 2009). 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’ to which it should apply. Definitions are given in the report *Principles and Guidelines for Access to Research Data from Public Funding* (OECD, 2007), which has strongly influenced UK Research Councils’ data access policies. These set out further principles to guide researchers and in some cases require funding bids to be accompanied by Data Management Plans. The OECD *Guidelines* define 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.”

Some limitations are seen as legitimate: this definition does not explicitly state that access must be public, nor without limits on its reuse. The *Guidelines* are also limited to certain kinds of data, excluding those gathered for commercialisation or private sector data, as well as datasets restricted for individual privacy, confidentiality, or for national security reasons. Current UK funding body data policies also typically allow researchers ‘first use’ of data, i.e. to embargo data until they have published from it.

The OECD *Guidelines* encourage authors to release datasets relevant to the claims made in their articles. Concepts of *open data* typically extend beyond this to include *pre-publication* data release. The *Guidelines* define data broadly as “factual records … used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings". They exclude physical objects and "laboratory notebooks, preliminary analyses, and drafts of scientific papers, plans for future research". The term '*Open Notebook Science*', (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."[2] Open notebook tools take the form of an electronic laboratory notebook (ELN), where the 'what happened' details of research are recorded alongside experimental data. The focus is on keeping a full record of experiments, serving as provenance information for published datasets.

The Science Commons *Open Science Principles*[3] also relate to a broad spectrum of research products, as do the *Panton Principles*[4], a set of recommendations for making scientific data open. Definitions carry various assumptions about what should be shared with whom, when, and how; what can legitimately be withheld; and what gains technology can bring to the equation. For some advocates 'open' refers to the absence of legal restriction on reuse, and for others social and technical aspects of accessibility and reusability are also relevant. To explore how these various aspects accord with practitioners' experience, we defined 'openness' broadly to describe how far research products are accessible and reusable beyond those contracted to produce them. This entails a continuum from proprietary control to limitless reusability.

Our literature review grouped the claimed advantages of working openly, and barriers to it, around five main issues: speed and efficiency of the research cycle; capabilities to identify new research questions; research effectiveness and quality; innovation, knowledge exchange and impact; and research group and career development.

Speed and Efficiency of the Research Cycle

Openness in science is credited with impacts on the speed and productivity of the research cycle. Yet few economic studies of research consider different forms of openness. According to David et al., (2009), economic studies of research productivity tend to dichotomise private and publicly-funded research, with the latter deemed "open" because the results are made public. Models indicate that the benefits of this openness, i.e. lower research costs of accessing prior knowledge, depend on public funding of post-graduate researchers to access it (Carayol & Matt, 2006; Mukherjee & Stern, 2009). However, it is less clear where these benefits stem from, i.e. how widely post-docs need to share data or methods, or how far in advance of publication.

Economic benefits can also be obtained from *public* participation in the research process (Lyon, 2009). According to Silvertown (2009), along with web and mobile-based tools that allow volunteers to gather (for example) field data on environmental change, one factor "driving the growth of citizen science" is "the increasing realisation among professional scientists that the public represent a free source of labour, skills, computational power and even finance."

[2] Open Notebook Science - Wikipedia. Retrieved January 19, 2010 from http://en.wikipedia.org/wiki/Open_Notebook_Science.
[3] Science Commons » Principles for open science. (n.d.). Retrieved March 30, 2010, from http://sciencecommons.org/resources/readingroom/principles-for-open-science/.
[4] Panton Principles. (2010). Retrieved February 19, 2010, from http://pantonprinciples.org/.

Studies identifying measurable economic benefits of open research are few; however, potential efficiency benefits are considered by Fry et al., (2008). They identify three areas of return on investment from secondary uses of data: reduced costs of collection and duplication; sharing direct and indirect costs of collection; and new uses unforeseen at the time of collection, including data mining opportunities. Each additional time a dataset is used/reused represents a direct financial benefit equivalent to the cost of collection. Indirect cost savings include more efficient use of scarce resources used in collecting data, including research subjects and instrumentation. The main costs of reuse include the storage costs faced by a repository, and the costs of preparing the data for curation and sharing. These preparation costs, documenting data to recognised metadata standards, are likely to be significant according to the cost modeling studies by Beagrie et al., (2010). Other economic advantages identified with openness by Fry et al., (2008) 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".

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 'collaboratorie'. Bos et al.'s study (2007) of scientific collaboratories identifies three main barriers to scientific research scaling beyond informal, one-to-one collaborations. Firstly, transferring knowledge requires specialist expertise and may be tacit. Secondly, research culture affords freedom to pursue high-risk ideas and resists corporate control. Thirdly, they point to difficulties of cross-institutional work crossing formal institutional boundaries, including IPR issues. Wide disciplinary variations in these barriers are likely, depending on the availability of community standards and acceptance of 'data' as a separate entity from the context of its production (Lyon et al., 2010). The related 'tacit knowledge' aspects of research will influence potential gains from sharing. 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. While some tacit knowledge may be documentable, it is normally taken-for-granted, learned through membership of a community, or drawn from presence in-situ (Collins, 2001).

Capabilities to Identify New Research Questions

An enhanced ability to identify research problems is associated with data sharing. The OECD *Guidelines* argue that sharing "reinforces open scientific inquiry", promotes new research and the testing of new or alternative hypotheses. However, it is unclear how far practitioners apply a scientific ethos of openness as grounds for sharing their data. David et al., (2008) describe open disclosure in scientific enquiry as an ethos "to which members of the academic research community generally subscribe, even though the individual behaviours may not always conform to its strictures". 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 sharing (Beagrie et al., 2010). Extensive sharing and curation of genomic and proteomic data has driven data policy, and made possible entire new data-based fields, such as functional genomics and systems biology. And yet even in the genomic and proteomic fields data sharing is far from universal (Piwowar & Chapman, 2008) and case studies indicate that life science data sharing is generally more restrained (Pryor, 2009).

On the other hand, the machine-readability of publicly available datasets has motivated researchers in life science and many other disciplines to explore the potential of semantic web technologies. The rise of “data-intensive” research (Hey et al., 2009) involves fundamental changes in how research questions get asked of data. Here, public access enables re-analysis of individual and linked datasets, and the potential for this is boosted where semantic web standards are applied to render the terms and relationships in machine-readable form (Coles & Frey, 2009).

Research Effectiveness and Quality

The open disclosure and peer review of research results has long been held to ensure effective validation. The implications of open data for this traditional model are uncertain, but may depend on broader participation in peer review. From the traditional standpoint, peer reviewing datasets amplifies existing concerns about the time costs of reviewing and the short supply of reviewers, and raises new ones about understanding the data (Ware, 2008). However, according to some open science advocates, broader public participation entails a radical shift in the peer review process, potentially including contributions from citizen-scientists (Stodden, 2010). Issues around who peer-reviews data quality, and how, are mostly keenly felt in fields where data is shared prior to publication. An RIN 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,” (RIN, 2008).

Given that citation frequency is used as a measure of research article quality, a correlation between citations and online availability of datasets upon which the articles are based suggests an association between public data release and perceived quality. There are indications of this in microarray studies for clinical trials (Piwowar et al., 2007). It is not yet clear if this is due to authors citing articles *because* they have access to underlying data. Since a large number of co-authors on an article is strongly correlated with its subsequent citation (Wuchty et al., 2007) the association with data sharing may be more indirect, for example because larger collaborative projects, whose publications have many co-authors, are relatively more motivated to make their data accessible online or are more likely to comply with funding body and publishers’ mandates.

Innovation, Knowledge Exchange and Impact

The relationship of scientific inputs to outputs is now accepted to be ‘non-linear’, i.e. research involves exchanges between researchers and private and public enterprises, often through informal networks (Martin & Tang, 2007). Innovation entails involvement of the ‘end-users’ of research outputs. The shaping of an innovation 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 & Hyysalo, 2008). 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; commercial involvement may affect researchers’ neutrality. On the other hand, where research is intended for commercialisation, openness may limit income from licensing the IP rights (David et al., 2004).

The term “citizen science” was coined to refer to public deliberation of the societal impact of science, driven partly by wider awareness of such impacts from research, and demands for more transparent governance (Irwin, 1995). Social scientists have mediated dialogue between scientists and public groups, using, for example, focus groups. These approaches have helped develop ‘open consent’ models, by negotiating the ethical barriers to human subjects’ data being shared and re-purposed in longitudinal clinical research without renewed consent (Haddow et al., 2005).

Ethical questions have also been raised about community participation in deciding the ownership of data collected with them, or from them, and rights to research results they have had a stake in producing. A notable example is pharmaceutical research whose data is based on ‘traditional’ medicinal knowledge. IPR in the research outputs has been used to share patent benefits with indigenous peoples whose local knowledge it is derived from (Vermeylan et al., 2008). Parallels with “citizen science” arise in how academic and private scientific institutions appropriate the knowledge produced by users and citizen scientists, and its value (Delfanti, 2010). The implication is that research benefits may be more equitably shared if ‘citizen scientists’ share the ownership of research results, or at least participate in deciding such questions, rather than academics assuming the responsibility of placing them in the public domain.

A common economic view of scientific impact on innovation is that cooperative open disclosure and competitive proprietary exchange 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 (e.g., David, 2004). The countervailing efforts of the Creative Commons project are significant not only for their ‘copyleft’ and public domain model licenses but also their moves to develop machine-readable licenses and ‘policy languages’ to reduce the human effort in dealing with multiple licensing models.

Research Group and Career Development

The need for more effective data citation and attribution mechanisms is key to career development. Although researchers who share their data may receive acknowledgements or direct citations to their datasets, the limited impact and recognition for data publication is one of the issues driving initiatives to standardise citation methods (Brase et al., 2009).

Economic studies of motivation for openness are limited to open source software development. Motives include ‘signalling of ability’, the need for a particular software solution, or mastering software challenge, and the desire of belonging to the ‘gift society’ of active OSS programmers (e.g., Bitzer, 2007). These factors suggest that social networking among early-career researchers might drive open publishing of research data and notebooks. One recent survey indicates very limited take-up of public web 2.0 platforms for scholarly communication, mostly confined to senior researchers. Junior and younger researchers are more likely to be frequent users of social networking (RIN, 2010). However the study found lack of clarity on benefits, and mistrust of sharing openly on platforms that lack standards for attributing effort.

Methodology

The project involved six case studies over a 14 week period, each based on semi-structured interviews with at least one established researcher at Principal Investigator (PI) level and, in most cases, others working earlier in their career. The interviews involved 18 participants working across six institutions. The disciplines involved were astronomy, bioinformatics, chemistry, epidemiology, language technology, and neuroimaging. To gather a broad range of views we sought out known advocates of open working, but also groups who we knew to be releasing data more selectively, including some we knew to be skeptical of the benefits.

The Science Commons 'Principles for Open Science' (op.cit.) were used as a starting point for the interviews, which were transcribed and summarised, and then fed back to each participant for comment. Summaries were then consolidated in a draft of the case study report, which was again fed back for comment. Additional desk research identified examples that participants had highlighted in interviews, or in publications and web resources relating to their group or project. The interviews used a topic guide which is available from the project website along with transcripts.

Open Advantages and Disadvantages: Issues for Researchers

The participating researchers were working openly to different degrees and with a variety of research resources. The case study report (RIN, 2010a) and its Appendices [5] elaborate on the following very brief outline:

- **Astronomy:** collaborators in the *Astrogrid* Virtual Observatory project had curated open data catalogues, and developed open metadata standards and data management software, some of which has been translated for medical application.
- **Bioinformatics:** members of an *Image Bioinformatics Research Group* had developed a software architecture for linked open data, applied to functional genomics and translated across several disciplines, and produced exemplars of semantically-enhanced journal articles with linked data.
- **Chemistry:** collaborators on *LabBlog*, a web-based lab notebook, had used it to aid compliance with lab safety regulations, share experimental records with colleagues or as public 'open notebooks', and linked to open repositories as source and destination of structured crystallography data.
- **Epidemiology:** researchers had extensively reused public health data, integrating this with both freshly collected and openly sourced geo-spatial data, and making results publicly available.
- **Language Technology Group:** members collaborated in various consortia producing multi-modal corpora for cross-disciplinary research in human interaction, and openly release much of this data and annotation tools.
- **Neuroimaging:** members of a research lab shared metadata publicly, imaging data on a more limited basis due to its sensitivity, and analysis tools with past and present collaborators, and pooled subject recruitment effort with collaborators.

[5] Project page: http://www.dcc.ac.uk/projects/open-science-case-studies.

Interviews themes were analysed according to the five main issues identified in the literature review, summarising pros and cons that participants raised. Examples were collected to demonstrate advantages claimed for working openly, and a typology to assess various dimensions of openness was developed. The main advantages were:

- **Efficiency in the research cycle:** saving data collection costs through reuse; indirect savings in research costs though use of pooled resources (avoiding recruitment fatigue); lower barriers to communication and collaboration; and lower cost barriers through open source 'gift exchange'. In some fields the greater efficiencies achievable in the research process had brought a step change "…the process of making things openly available so that derivative science can be done is massively speeded up, so all in all the wheels turn quicker," (Chemical Crystallography, PI).
- **New research capabilities:** As in the example above, new capabilities to find research questions and analyse evidence were being realised. In the language technology case, fields of research were made feasible through building cheaper or better models from open source components. Other cases included data mining from open repositories, or new abilities to find patterns through visualising across previously disparate data sources.
- **Effectiveness:** more potential for scrutiny, and for cross-disciplinary collaboration. Compliance with regulatory scrutiny requirements has spin offs in some areas: "…because you have to put the effort in anyway…the data that you collect later on already has much higher quality meta data, or could have, associated with it without extra effort,"(Chemistry, PI).
- **Knowledge exchange and impact:** commercialisation opportunities, and higher visibility for researchers and institutions were generally motivators and had produced results for some: "…we have a federation of all sorts of different types of repositories here, and it's perceived very strongly from the institution from all sorts of different angles, you know, you can use it for administrative processes, you can use it for tying into continued professional development and promotion, and the increased visibility of the institution out there," (Chemical Crystallography).

Participants did not, however, see openness as a binary choice: "Degrees of openness is extremely important. Different people work in different ways and have different constraints imposed upon them," (Chemistry, Senior Researcher). Like Fry et al., (2009) we found it helpful to consider two main 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.

In the study report we combine these dimensions to provide a matrix of examples, to aid comparison across cases. The first of the dimensions can be used to describe material outputs of each step in the research cycle, as 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, FAQs, supplementary information
Publishing and reporting	Conference papers, journal articles, technical reports
Engaging and translating	General articles, web pages, briefings, public exhibits, presentations
Infrastructuring	Software tools, databases, repositories, web services, schemas and standards

Table 1. Research Cycle Stages and Material Outputs.

The second dimension consists of six main ‘degrees of openness’; a continuum of research materials disclosure by creators to other actors in their production and potential reuse. It begins with actors formally (contractually) tied to the creators and extends to those not previously linked to them, i.e. the general public. We characterise these degrees of openness as follows:

- **Private management:** sharing within a research group, where resources are organised to facilitate access and reuse by researchers within the group, to include at least some data or metadata on all research activity.
- **Collaborative sharing:** sharing between members of a consortium established to deliver a project or programme, so that researchers employed may access data or metadata, e.g., on an intranet, and reuse it for their common contractual purpose.
- **Peer exchange:** sharing on the understanding that disclosure or reuse have conditions attached, between members of a researchers’ network of peers, e.g., using a social networking web platform.
- **Transparent governance:** disclosure to an external party according to a publicly accountable code e.g., enforced by institutions or funders for research assessment, ethical scrutiny, or safety inspection; or where sharing is facilitated by a third party such as an archive with an institutional or funding body mandate.
- **Community sharing:** access or reuse limited to identifiable members of a research community or communities, e.g., defined by affiliation to an institution, research network or association; facilitated by collaboratories or virtual research environments, and resources licensed for educational access/use.
- **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 reused by a designated research community, and with few restrictions - such as a limited embargo period.

Issues affecting the feasibility and desirability of open working to these varying degrees (which are not mutually exclusive) are discussed below.

– *Public good or source of competitive advantage?*

"…We do have a few projects of some industrial importance, for which we are obliged to be closed," (Chemistry, PI). All researchers saw an obligation to disclose data produced using scarce public resources (instrumentation, software tools, unique observations), especially where there was scientific justification e.g., to better enable comparison of results, cross-disciplinary collaboration on analysis techniques being a common driver. Limited embargos to allow results to be produced were thought legitimate. Openness inhibited some commercial collaboration, but also offered greater visibility to research groups and institutions, and opportunities that could be pursued.

– *How will disclosure benefit the research?*

The relative size and cohesion of the researchers' communities were a factor in judging whether or not it was worth disclosing. In some cases, researchers felt they knew everyone working in their specialist field and believed that making their data understandable beyond that, except for occasional public engagement purposes, would damage their productivity and/or career progression, e.g., "…if I had to then make it polished so that people could follow what I'm doing, I'm sure there wouldn't be many people interested in following what I was doing …if I spent a lot of time trying to make it understandable for people outside my field but I don't think it would. I think I'd spend an awful lot of time doing it," (Chemistry, Post-doc).

– *How to justify data-linking infrastructure?*

Funding body support for research groups to invest in this was considered an exception. Success was believed to be judged on scientific merit, i.e. novel peer-reviewed results, rather than on criteria more appropriate for infrastructure. More coordinated provision of training in research data management was also called for. This human infrastructure can be critical: "…for larger projects it's very easy though, we'll have one or two dedicated people dedicated to this aspect of the data and so they can become trained in how to do this. I think it's much slower for smaller research teams; it will take a long time before they routinely put their data into the virtual observatory," (Astronomy, PI).

– *How feasible is documentation and quality assurance for reuse?*

There were concerns about the practicalities of going beyond the current depth of detail or breadth of disclosure, owing to lack of skill, accepted process, or capabilities needed for a wider audience to judge its quality and avoid misinterpretation: "...if you flood the literature with lots of stuff that isn't very good then it's a problem... now putting out the raw data… it's really necessary for the considered stuff supporting your publication but I'm a little more wary of doing it generally. Publishing everything raises quality control and interpretation issues," (Chemistry).

– *How to justify packaging and metadata creation costs (and anonymisation)?*

Resourcing the effort for these is an issue. Resources for end-of-project 'packaging', e.g., presenting data according to repository ingest procedures or documenting software, are sacrificed through time constraints or given a lower priority than the initial analysis or development. In neuroimaging, for example, anonymisation issues make public data sharing rare. Code sharing is relatively common and analysis software is rapidly developing, but often depends on local knowledge, leading the neuroimaging participants to share code with a network of trusted peers "…we're continuing to work with these groups to in effect have a shared environment, which is very warm share because they're people we have worked with and they understand what our thinking is," (Neuroimaging, Senior Researcher).

– *How to justify a choice of standards to adopt*?

Some participants had been involved in developing standards for data and metadata, but recognised that take-up was inhibited by fear of 'the wrong choice' leading to adverse impacts on practice: "… if they had the suspicion that it's the latest technology fad and in five years time we'll all be asking them to do something completely different, and that will have gone, then they'll have wasted their time trying to do things a certain way," (Astronomy, PI).

Conclusions: Openness by Degrees

The case studies presented here capture a range of views or practices, and they show that even skeptics see advantages in opening up 'just enough' of their working practices for pragmatic ends. However, the evidence-base to support claims made for open working is still under-developed. Our studies were exploratory and we do not claim that our participants are representative. But recent surveys show only a small minority of researchers using open methods, and there is ongoing debate on what 'open science' should encapsulate.

Many of the benefits envisaged for open methods relate to how far they enable not only access but active participation in a research community by newcomers and outsiders, and maintain low barriers to this participation. This entails decisions for policy makers, research investigators and user communities on the risks, costs and benefits of broader participation before and after results are produced. Greater visibility for research producers, lower barriers to collaboration, and more reusable datasets are strong motivations. However, there is a need to be able to identify the costs of packaging data and describing it to be reusable for broader purposes. The defining 'value proposition' for openness could itself be tested against the economic criterion of a public good, i.e. sharing should increase rather than decrease its value. The criteria and measures used by archives in data appraisal offer broad parameters for deciding what is worth sharing. However, like Cragin et al., (2010) we see a need for further research and guidelines to help researchers identify the risks and benefits associated with data types and practices specific to their communities, in order to better prepare for opening up the research process.

We aim to explore further how research communities' expectations of reciprocity, and means of reciprocation, affect patterns of disclosure. There is also a need for more concrete evidence of the benefits gained as research communities find ways to be more open at each stage of their research lifecycle. To support them we also aim to develop the openness typology as a template for providing clearer guidance on the planning and decision making required for more open research.

Acknowledgements

The authors wish to express their great appreciation to all the researchers who gave their valuable time to participate in the interviews. We also thank reviewers of the paper for their very helpful comments.

References

Beagrie, N., Lavoie, B., & Woollard, M. (2010). *Keeping research data safe (Phase 2)*. Retrieved February 10, 2011, from http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx.

Bitzer, J., Schrettl, W., & Schröder, P. (2007). Intrinsic motivation in open source software development. *Journal of Comparative Economics*. Retrieved February 10, 2011, from http://linkinghub.elsevier.com/retrieve/pii/S0147596706000643.

Blumenthal, D., Campbell, E., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S., & Holtzman, N. (2006). Data withholding in genetics and the other life sciences: Prevalences and predictors. *Academic Medicine*, *81, (2).*

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).*

Bradley, J., Owens, K., & Williams, A. (2008). Chemistry crowdsourcing and open notebook science. *Nature Precedings*. Retrieved February 10, 2011, from http://precedings.nature.com/documents/1505/version/1.

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).*

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

Coles, S., & Frey, J. (2009). *The relevance of linking*. Retrieved February 10, 2011, from http://ie-repository.jisc.ac.uk/419/.

Collins, H.M. (2001). 'What is tacit knowledge'. In T.R Schatzki, K. Knorr-Cetina & E. Von Savigny (Eds) *The practice turn in contemporary theory*. Routledge.

Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science, and institutional repositories. *Philosophical Transactions of the Royal Society A*, *368, (1926).*

David, P.A. (2004). Can “Open Science” be protected from the evolving regime of IPR protections? *Journal of Institutional and Theoretical Economics JITE, 160*. doi:10.1628/0932456047738610069.

David, P.A., den Besten, M.D., & Schroeder, R. (2008). Will e-Science be open science? *SSRN eLibrary*. Retrieved February 10, 2011, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1317390.

David, P. A., den Besten, M.D., & Schroeder, R. (2009). *Collaborative research in e-Science and open access to information.* SIEPR Discussion Paper No. 08-21. Stanford University.

Delfanti, A. (2010). Users and peers: From citizen science to P2P science. *JCOM: Journal of Science Communication*, 9. Retrieved February 24, 2011, from http://jcom.sissa.it/archive/09/01/Jcom0901%282010%29E/.

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J., & Rasmussen, B. (2008). *Identifying the benefits of curating & sharing research data*. Retrieved February 10, 2011, from http://www.jisc.ac.uk/publications/reports/2008/databenefitsfinalreport.aspx.

Haddow, G., Cunningham-Burley, S., Bruce, A., & Parry, S. (2005). *Generation Scotland preliminary consultation exercise 2003-04: Public and stakeholder views from focus groups and interviews*. Institute for the Study of Science, Technology and Innovation, University of Edinburgh.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

Irwin, A. (1995). *Citizen science: a study of people, expertise, and sustainable development*. New York: Routledge.

Lyon, L. (2009). *Open science at web-scale: Optimising participation and predictive potential.* Consultative Report. Retrieved July 21, 2010, from http://www.jisc.ac.uk/publications/reports/2009/opensciencerpt.aspx.

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 February 10, 2011, from http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf.

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

Meagher, L., & Lyall, C. (2009). *The invisible made visible: The role of evaluation in informing processes of knowledge exchange*. Retrieved February 10, 2011, from http://www.genomicsnetwork.ac.uk/innogen/publications/workingpapers/title,21156,en.html.

Mukherjee, A., & Stern, S. (2009). Disclosure or secrecy? The dynamics of open science. *International Journal of Industrial Organization*, *27, (3).* doi:10.1016/j.ijindorg.2008.11.005.

Nature (2009) Special issue on data sharing. *Nature.com.* Retrieved February 10, 2011, from http://www.nature.com/news/specials/datasharing/index.html.

OECD (2007). *Principles and guidelines for access to research data from public funding.* Retrieved February 10, 2011, from http://www.oecd.org/dataoecd/9/61/38500813.pdf.

Piwowar, H., & Chapman, W. (2008). Prevalence and patterns of microarray data sharing. *Nature Precedings*.

Piwowar, H., Day, R., & Fridsma, D. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, *2, (3).* http://dx.doi.org/10.1371/journal.pone.0000308.

Pryor, G. (2009). Multi-scale data sharing in the life sciences: some lessons for policy makers. *International Journal of Digital Curation*, *4, (3)*.

RIN. (2008). *To share or not to share: Publication and quality assurance of research data outputs*. Retrieved February 10, 2011, from http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs.

RIN. (2010) *If you build it, will they come? How researchers perceive and use Web 2.0.* Research Information Network. Retrieved February 10, 2011, from http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/use-and-relevance-web-20-researchers.

RIN. (2010a) *Open to all? Case studies of openness in research*. Retrieved February 10, 2011, from http://www.rin.ac.uk/our-work/data-management-and-curation/open-science-case-studies.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24, (9).*

Stewart, J., & Hyysalo, S. (2008). Intermediaries, users and social learning in technological innovation. *International Journal of Innovation Management*, *12, (3).*

Stodden, V. (2010). Open science: policy implications for the evolving phenomenon of user-led scientific innovation. *JCOM*, *9, (01).* Retrieved February 10, 2011, from http://jcom.sissa.it/archive/09/01/Jcom0901(2010)A05.

Vermeylen, S., Martin, G., & Clift, R. (2008). Intellectual property rights systems and the assemblage of local knowledge systems. *International Journal of Cultural Property*, *15, (02)*.

Ware, M. (2008). Peer review: Benefits, perceptions and alternatives. Retrieved February 10, 2011, from http://www.publishingresearch.net/documents/PRCsummary4Warefinal.pdf.

Wuchty, S., Jones, B.F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316, (5827).*