Lessons for data sharing from institutional repositories

Rebecca Parker¹, Dana McKay², Terrence Bennett³

¹,² Library and Institute for Social Research, Swinburne University of Technology, Hawthorn, Australia, rparker@swin.edu.au, dmckay@swin.edu.au
³ TCNJ Library, The College of New Jersey, Ewing NJ, United States, tbennett@tcnj.edu

INTRODUCTION

Increasingly, governments and the public are calling for research data to be publicly accessible; this is partly tied to the perception that this data is paid for with public money and therefore should be a public good [1]. In a time of contentious issues such as climate change, accountability also weighs into this debate—when death threats are being made against climate scientists, scientific transparency and public input are vital to a continued government mandate for research funding [2]. These changes in science policy have the support of many governments; in Australia the Government 2.0 Taskforce, the Australian National Data Service (ANDS) and numerous government data websites are specifically targeted to this purpose. Similarly there are websites devoted to sharing government data in New Zealand, the UK, the US and beyond. Research funders also support public access to data: in Australia the Australian Research Council and the National Health and Medical Research Council in conjunction with Universities Australia produced the Australian Code for the Responsible Conduct of Research, which mandates research data should be available beyond the life of any research project, via an institutional repository ‘where possible’. The interest of government and compliance with the Code are key motivators for universities in encouraging researchers to share data; evidence of adherence to the Code is instrumental in securing research funding. In addition, there is long term benefit to institutions in having greater awareness of their own research activity, and in showcasing the best of it via institutional mechanisms.

While institutional and government impetus for data sharing and curation are clear, it is individual researchers who will share— or not share—research data. While data sharing is arguably part of a scientific tradition [3], and some have argued that researchers with an objection to sharing data are being antiscientific [4], neither threats of enforcement of the Code nor these existing traditions is enough to encourage researchers who are not part of a data sharing tradition to share openly. The twin behaviours of reticence and rejection of institutional solutions mirror the problems faced recruiting publications to institutional repositories; thus we use institutional repository publication recruitment as a model for understanding the problems facing data curation. In this paper we will discuss five of those problems—(dis) incentives, difficulty, danger, demands and disciplinary data sharers—with reference to the literature on institutional repositories, before drawing conclusions on the lessons repositories can offer the open data movement.

(DIS)INCENTIVES?

One of the major reasons given for not depositing publications in an institutional repository is lack of incentive to do so, even in the face of institutional mandates and some suggestions that open access to publications improves citation rates (for example [5]). In light of this, it is perhaps unsurprising that researchers are not rushing to share their data; the Code does not specify public gain or promotional activity as the reason for sharing data, but instead is focused on ‘justifying’ the outcomes of the research and ‘defending’ them if they are challenged. This message is immediately discomfiting to researchers, who already feel that they operate in a climate of constant research performance measurement, and some of whom believe others would not understand their data [6]. Disappointingly, we cannot even promise researchers the increased citation benefits of open access publications when data sharing—while one study [7] has made a connection between shared research data and increased citation impact for the associated publications, the study was restricted to a single narrow research area and the results do not seem to have been replicated in other disciplines. While direct data citation models have been proposed, academia has not readily adopted the practice—more commonly, researchers mention the use of others’ research data in the acknowledgements section of papers, making it impossible to track the usage of research datasets as we do publications.

DIFFICULTY

Difficulty depositing publications has always impeded content recruitment in institutional repositories, and researchers sometimes see a requirement to deposit as another administrative burden imposed by universities [8]. Since the argument that open access data is personally beneficial to researchers is even weaker than for open access publications, the deposit process should be much easier for data. Unfortunately the converse is the case. Publications are described in databases by standard metadata; this is not universally true of research data. One solution to difficulty in the publications world is ‘mediated’ deposit, where repository librarians are responsible for record creation and copyright clearance [9], however this model does not translate well to data. In a trial at Swinburne, we discovered that data deposit requires input from the academic, however academics’ difficulties with metadata creation are already well known [9]. The bar for creating good

⁴ DataCite: http://www.datacite.org/
quality data records is very high for researchers and librarians alike; given the low likelihood of personal benefit it is easy to see how researchers find any data deposit obligation burdensome.

**DANGER**

Along with difficulty, the risk to researchers’ reputation in having work in an institutional repository is (wrongly or rightly) a major disincentive to deposit [8]. These dangers are even more pronounced for data: publications repositories have already had to address questions of quality control, provenance and policy, including researchers’ uncertainty about their contracts with scholarly publishers. In the data context, however, researchers believe they risk breaching the privacy of their participants, whose needs are more concrete to them than the abstract requirements of a faceless funding body. They are also concerned that many of those who access their data will be ill-equipped to understand it, and that the data may not be treated with respect, especially by those outside their field [6, 8]. These reputational risks leave many researchers unwilling to share data even within their own field [10], and even if it is a condition of publication [7, 11]. Given these concerns, many researchers are a long way from sharing their data freely and openly.

**DEMANDS**

The institutional repository literature supports mandating deposit of research in repositories. Yet even when mandates are in place, there is still a relatively low compliance rate [12], which can again be attributed to lack of incentive, despite the possibility of increased citation. There is no structure in place for rewarding research data contributions, so attempting to enforce compliance is even less likely to be successful. Even those who believe that research data sharing is a ‘collegial tradition’ warn that this tradition fails when the terms and nature of data sharing are dictated by a third party [3].

**DATA SHARERS**

Contrary to the difficulties outlined above, some researchers are already freely sharing their data because data sharing is part of their disciplinary research culture [6]. This activity mirrors the well-established sharing of publications through subject repositories; in some disciplines both data and publications are shared [6, 8]. While disciplines where data sharing is well established may seem ideal candidates for data curation via institutional repositories (after all, danger and difficulty—potentially the largest barriers to sharing—have already been managed) when we look to the repository literature, we see that those who used disciplinary repositories were disinclined to contribute to their universities’ repositories [8]. Institutional repositories were perceived as second rate and as not providing the same benefits as disciplinary repositories [8]; there appear to be similar attitudes with respect to data on the part of data sharers [6].

**CONCLUSIONS**

If data curation policies are to be enforced, all five of the major barriers to data management must be overcome, and while difficulty and danger are large barriers, disinclination alone is enough to prevent data deposit. Moreover, for data repositories to meet the aims of institutions and national bodies (gaining recognition for high quality work and retaining the fruits of public investment in research), data deposit must be near-universal. Ignoring or alienating either those with sharing cultures already in existence, or the more reticent data sharers, is contrary to these aims.

Given researchers’ real and valid concerns, and the scope of the problems with data curation, it is important that neither individual researchers nor institutions are penalised, at least in the initial phase of government-driven data curation activity, for not complying with a policy which is at least at this stage unenforceable. The lessons from institutional repositories can be used to point the way to strategies for making data sharing more palatable to researchers, and therefore encourage willing compliance with data policy.

**REFERENCES**

ABOUT THE AUTHORS

Rebecca Parker graduated from Curtin with a Master of Information Management in 2007, and is currently the Research Services Librarian at Swinburne University of Technology. She has managed Swinburne’s institutional repository since 2008, and more recently is involved in planning Swinburne’s support for research data management, supporting online publishing, and developing the content for researcher profile pages. Rebecca is keen to ensure that academic library services are tailored not just to the needs of students, but also to researchers and corporate areas of universities. She served as subject expert on the ARROW-funded NicNames Project, which explored how best to display name variants in digital libraries from a researcher perspective. She was a member of the ARDC Party Infrastructure Project Advisory Group, a committee set up by the National Library of Australia in partnership with the Australian National Data Service. Rebecca is also a regional coordinator for the ALIA New Graduates Group and is passionate about bringing new people and perspectives to libraries.

Dana McKay comes from an academic background in computer science, earning a Masters degree in digital library usability at the University of Waikato in 2001. Following her degree she worked as a research fellow and usability analyst at the University of Waikato focusing on information seeking and retrieval, and as a usability analyst at Nokia focussing on mobile usability. Since 2007 she has been working at Swinburne University of Technology Library, investigating a range of issues for user experience and information seeking, including author names, institutional repository usability, and researchers’ approach to research data management.

Terrence Bennett is currently the Business and Economics Librarian at The College of New Jersey; prior to that, he held a similar position at Emory University. His current research is focussed on data services in academic libraries, and in 2010-2011 he worked as the Research Data Librarian at Swinburne University of Technology, in support of a project funded by the Australian National Data Service. Terrence has a Master of Science in Library Science from the Graduate School of Library & Information Science at the University of Illinois, where he is currently an adjunct instructor of Business Information. He has given numerous presentations on research data services, information literacy, and business information resources for IASSIST (International Association for Social Science Information Services and Technology), the Association of College and Research Libraries, and the Special Libraries Association. Terrence is also a former co-chair of the American Library Association’s PRIMO (Peer-Reviewed Information Materials Online) committee.