The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

The Informatics Transform: Re-Engineering Libraries for the Data Decade

Liz Lyon,
UKOLN

Abstract

In this paper, Liz Lyon explores how libraries can re-shape to better reflect the requirements and challenges of today's data-centric research landscape. The Informatics Transform presents five assertions as potential pathways to change, which will help libraries to re-position, re-profile, and re-structure to better address research data management challenges. The paper deconstructs the institutional research lifecycle and describes a portfolio of ten data support services which libraries can deliver to support the research lifecycle phases. Institutional roles and responsibilities for research data management are also unpacked, building on the framework from the earlier Dealing with Data Report. Finally, the paper examines critical capacity and capability challenges and proposes some innovative steps to addressing the significant skills gaps.

International Journal of Digital Curation (2012), *7*(1), 126–138. http://dx.doi.org/10.2218/ijdc.v7i1.220

Introduction

Research libraries have traditionally supported the scholarly research and communication process, largely through supporting access to and preservation of its published outputs. The library cornerstones have been positioned around a long-established publication process tailored to deliver the peer-reviewed scholarly article or monograph; but now the research landscape is dramatically changing. The application of computational science and growth of data-intensive research, combined with a veritable explosion of social media tools and Web technologies, are reshaping research practice. Data in diverse formats are at the heart of the research process, but there are significant gaps in infrastructure to effectively share, manage, curate, preserve and potentially reuse the rapidly growing volumes of data generated by text corpora, video collections, 'big science' endeavours and by the expanding long-tail of 'small science'.

The Informatics Transform Proposition

This infrastructure encompasses hardware and software components for data integration, manipulation, recombination and storage, but also includes the essential human infrastructure required at an institutional level to advise, guide, train, co-ordinate and lead the stewardship effort to develop data management capacity and capability. Academic libraries have performed a similar stewardship role in the past; now is the time to critically examine their structure, function and service portfolio, to ensure that they are fit for purpose to support the data-intensive research of today effectively. The 'deconstruction' process described in this article, draws on the principle underpinning the mathematical analysis known as the Fourier Transform.[1]

Research in the Data Decade

Data scale and complexity has attracted many metaphors: described in 2003 by Tony Hey and Anne Trefethen as the 'data deluge' (2003), as 'big data' (Nature, 2008), 'surfing the tsunami' (Pain, 2011) and by the data-intensive scientific discovery in the *Fourth Paradigm* (Hey, Tansley & Tolle [Eds], 2009). In this collection of essays, the late Jim Gray described the emerging evolution of complementary arms of a discipline: Comp(utational)-X and X-Info(rmatics). The latter encapsulates the collection and coding of data and information which are required to facilitate the effective scaling of data-intensive research programmes. Implications for research stakeholders including institutions, were described in the *Dealing with Data Report* (Lyon, 2007) and were presented as a set of roles, rights, responsibilities and relationships. We will revisit them below.

[1] In 1822, the French mathematician and physicist Jean-Baptiste Joseph Fourier published his observations on heat propagation, making a number of mathematical assertions. This original work subsequently informed the derivation of a key mathematical analysis which has become known as Fourier Analysis. In simple terms, this is the decomposition of a function into its constituent parts; the *Fourier Transform* describes the deconstruction process.

 doi:10.2218/ijdc.v7i1.220

In parallel, an increasingly open scholarly communications agenda which promotes the enhanced sharing of research data outputs, has emerged (Borgman, 2011), fuelled by the adoption of Web tools and applications which accelerate the 'publication' process, and by economic drivers for greater accountability and transparency, to show the impact of public investments in science. Research infrastructure in institutions and academic libraries was explored in the *Open Science at Web Scale Report* (Lyon, 2009), data issues were raised but there appeared to be little appetite for change .

Then in April 2011, the EPSRC published its Policy Framework on Research Data, joining all UK research councils (and other global research funding bodies e.g. US National Science Foundation), in setting clear guidelines and expectations for Higher Education institutions and grant holders for proactive research data management planning and good practice. This policy steer has in a sense, provided validation for the *Informatics Transform,* which seeks to reshape libraries in a data-centric world.

Implementing the Informatics Transform: Some Assertions

So how should research libraries, library staff, librarianship (and information science), re-position, re-profile and re-structure to be fit for purpose in a data-centric research landscape? Here are five assertions as potential pathways to change.

Data Informatics and the University Library: A Catalyst for Institutional Research Data Management?

The research data landscape can be deconstructed in several ways, each of which produces a valuable perspective. Firstly we will examine the institutional research data lifecycle.

The University of Bath has recently been awarded JISC funding to implement the Research360 Project[2] (R360), to embed good research data management practice within the institution. R360 develops the concept of an end-to-end 360-degree institutional research lifecycle which covers research conception, provision of good practice advocacy, supports the production of data management plans and associated training, the role of doctoral training centres (DTCs), storage of research proposals in a Research Information Management (RIM) or Current Research Information System (CRIS), to generating or collecting primary data in research workflows, structured data curation and deposit in an institutional data repository or file-store, through to development of institutional policy, data publication and reuse, with associated credit and attribution (see Figure 1 below).

In this scenario, the Director of Information Services (or Chief Information Officer or University Librarian), is positioned to contribute authoritative informatics leadership and co-ordination working in partnership with the PVC Research, to assure clarity of vision supported by an understanding of the research data landscape. In particular, changing policy requirements and their implications for HEIs should inform the development of institutional research data management policy.

[2] Research360 blog: http://blogs.bath.ac.uk/research360/

But how do we ensure that this informatics leadership is in place? One answer is for professional organisations (SCONUL, Research Libraries UK and CILIP), to accept a collective responsibility to ensure that Library 'leaders-in-waiting' are given the appropriate leadership training to equip them to operate in this data-centric world. It is perhaps worth noting a rather controversial point here, that there is a potential benefit for those leaders to have been practising researchers at some point during their career, whether in physical sciences, social sciences or the arts and humanities. The domain is not critical; the key requirement is a first-hand understanding of the research process and its associated data informatics requirements.

In Figure 1, REF is the UK Research Excellence Framework, a system for assessing the quality of research; RCUK is the strategic partnership of the seven UK Research Councils.

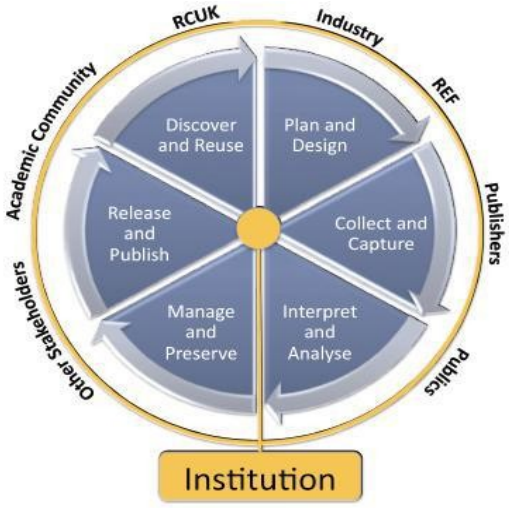

Figure 1. The Research360 Institutional Lifecycle Research Concept.

We can further deconstruct data informatics by unpacking library support services for research data management (RDM). These might include:

1. **RDM Requirements**: supporting departments to carry out research data requirements surveys (which might address both legacy and future data needs), using data audit and assessment tools such as the *Data Audit Framework (DAF)* and *CARDIO* from the Digital Curation Centre (DCC).[3]
2. **RDM Planning**: working with staff in faculty Doctoral Training Centres to deliver advocacy and guidance to post-graduates, research staff and PIs, on effective data management planning, using tools such as the UK Digital Curation Centre *DMPOnline* tool or the *DMPTool* from the California Digital Library. There is also an opportunity to contribute to institutional annual planning rounds, through probing questions which explore future data infrastructure demands, storage requirements, security and sensitivity issues, and access.

[3] Digital Curation Centre: http://www.dcc.ac.uk/

 doi:10.2218/ijdc.v7i1.220

3. **RDM Informatics**: providing technical expertise for structured data description, including use of metadata standards and schema, data formats, ontologies appropriate to particular disciplines and domains. This is an opportunity to promote disciplinary norms and to work towards consensus.
4. **RDM Citation**: supplying guidance and links to third-party services such as *DataCite,* to enable the assignment of persistent identifiers to datasets to support discovery, citation and reuse.
5. **RDM Training**: collaborating with faculty Doctoral Training Centres to deliver data management training programmes and modules to postgraduates, research staff and PIs. The DCC has developed data training materials (*DCC 101 Lite*) and there are a growing number of resources from JISC projects (MANTRA, CAIRO, DATUM, DataTrain, DMTpsych) that can be re-purposed.
6. **RDM Licensing**: pointing to expert guidance (e.g. *DCC Guide How to License Research Data*), to assist with queries about data licensing. This expert guidance may also cover legal and ethical issues associated with datasets. Engagement with university research ethics committees can be a valuable channel for communication and dissemination.
7. **RDM Appraisal**: providing guidance to assist with queries about which data to keep. The DCC has produced a helpful guide on *How to Appraise and Select Research Data.*
8. **RDM Storage**: collaborating with IT Services to ensure clarity and relevance of local data storage guidelines and infrastructure provision. There also needs to be a proactive approach to collaborating with disciplinary, national and international data centres and an awareness of funder-policy requirements (e.g. ESRC) for data deposit in such archives.
9. **RDM Access**: informing the development of pragmatic data release policies and embargoes, which support open science agendas and FOI, but also respect the intellectual property of the researcher, the institution and any collaborating industrial partners (the Research360 project has a particular focus in this area). Ensuring appropriate links from underlying and supplementary data to articles in institutional repositories.
10. **RDM Impact**: linking with Research Support Offices to collect impact evidence directly relating to research datasets and their reuse, for inclusion in RIM/CRIS systems. There are a number of emerging-impact tracking services such as *total.impact.org*, which facilitate the monitoring of collections of research objects including datasets. Libraries can promote their use and help researchers to maximise their 'network impact' on the Web.

Of course, delivery of these new data informatics services may require a review of library and information services organisational structure and resource reallocation. It is worth making an obvious point that each HEI is different with a unique local service permutation, and what works for one, will not work for all, but there are some common principles for success:

- **Adopt a partnership approach** which is open and inclusive e.g. follow a consultative track to RDM policy development with faculty and service staff.
- **Build co-ordination and coherence** in RDM planning across departments and services, e.g. ensure a common understanding of the external drivers, using shared templates and planning tools.
- **Facilitate service integration** across internal suppliers e.g. between the data management planning activity and the data storage provider. In this context, service integration does not mean converging services or continuing to run converged services or de-converging services; rather it implies implementing integrated services across multiple institutional stakeholders with joint planning and shared service development.
- **Ensure shared learning** across the institution and beyond e.g. provision of infrastructure to share RDM plans, RDM training modules as open educational resources.

We can also deconstruct research data management in terms of the institutional roles and responsibilities. Table 1 builds on the framework in the earlier *Dealing with Data Report,* but focuses on the Library and its key stakeholder relationships, within a research-led, Higher Education institution.

Role	Responsibilities	Requirements	Relationships
Director Information Services / CIO University Librarian	To lead and co-ordinate data informatics support.	Appropriate LIS structure in place. Library staff with data informatics & research data management skills. Institutional repository with content links to underlying research data.	PVC Research, Deans, Associate Deans, Faculty/School Directors of Research, IT Director, Director Research Support. Other key institutional stakeholders. Open Access Publishers.
Data librarian / Data scientist / Liaison /Subject / Faculty Librarian	To deliver expert data informatics advice and guidance to research staff. To facilitate access to datasets for PIs, research staff, postgraduate and undergraduate students.	Knowledge of data management planning and data audit and assessment tools. Knowledge of selection and appraisal, metadata standards and schema, data formats, domain ontologies, identifiers, data citation, data licensing. Knowledge of appropriate disciplinary data centres.	Doctoral training centres (DTCs), post-grads, PIs. DCC. DataCite. Data centre staff.
Repository managers	To ensure research papers have persistent links to underlying research data.	Knowledge of persistent identification mechanisms and publisher requirements.	Data librarians / Data scientists / Liaison /Subject / Faculty Librarians.

 doi:10.2218/ijdc.v7i1.220

Role	Responsibilities	Requirements	Relationships
IT / Computing Services	To provide data storage infrastructure and guidance.	Knowledge of data storage options including cloud-based services.	EduServ data centre. Cloud service providers. National data centres.
Research & Development Support Office / Research & Innovation Services	To provide RIM/CRIS capability for research outputs.	Provision for non-textual outputs such as datasets, software and program code, gene sequences, models.	Research funding bodies. Data scientists / Liaison /Subject / Faculty Librarians.
Faculty Doctoral Training Centres	To supply training to new-entrant researchers and PIs.	Knowledge of data management planning and data audit and assessment tools. Training programmes and modules.	Deans & Associate Deans, PIs. Data librarian / Data scientist / Liaison / Subject / Faculty Librarians.
PVC Research	To develop institutional research policy and code of practice.	Understanding of data management compliance implications, risks including legal and ethical issues, and sustainability challenges.	Deans & Associate Deans. Key service directors. Research & Development Support Office / Research & Innovation Services.

Table 1. Research data management, the library and institutional stakeholders.

Mainstreaming Data Librarians and Data Scientists?

The RDM services described will require specific informatics skills and knowledge associated with particular disciplines and domains. These might include a working knowledge of the research practices and workflows, an understanding of the specific technical standards, metadata schema and vocabularies routinely used in practice, an awareness of the national and international data centres where research data in that domain are deposited, and a good grasp of the data publication requirements of the leading scholarly journals.

Who provides this support today? Only a few UK university libraries have a designated data librarian e.g. Edinburgh. The University of Bath has recently appointed an institutional data scientist as part of the Research360 Project, to help to promote good research data management practice across the institution.

In contrast, most university libraries have a staff team (variously called liaison librarians, subject librarians or faculty librarians), whose primary role is to engage and support research and teaching staff and students in academic departments. Sometimes the liaison librarian is qualified in that particular discipline; interestingly, for information scientists in pharmaceutical companies, this may be an essential attribute. So can this library support team adapt to augment the data informatics roles embraced by data librarians and data scientists?

Strong informatics skills are key to delivering RDM services, and individuals with these skills will be in increasing demand. In a recent article about genomics and the torrent of DNA data (Pennisi, 2011), it was noted that bioinformaticists are in short supply everywhere and computational biologist Chris Ponting, University of Oxford, said '*I worry that there won't be enough people around to do the analysis*'. Is there a

similar shortage of librarians and information scientists equipped with data informatics skills? How can Library and Information Science (LIS) schools meet this capacity and capability shortfall? Next we shall consider LIS curriculum and career incentives.

Data Informatics Embedded in the LIS and iSchool Curriculum?

In recent years, iSchools in the United States have launched a flurry of new Masters courses with a major data curation component e.g. University of North Carolina Chapel Hill, at Illinois Urbana-Champaign and at the University of Michigan. The first UK iSchool, located at the University of Sheffield, is planning to offer a new undergraduate BSc in Informatics as well as domain-based informatics courses. More recently, the International Curation Education (ICE) Forum[4] in London during June 2011 provided an opportunity for discussion and development of the digital curation training curriculum. The British Association for Information and Library Education and Research (BAILER) currently has 20 members providing LIS education in the UK. How well do curricula reflect the requirements of a data-intensive research landscape, with elements of data informatics, research data management and data curation?

The Research Information Network (RIN) Working Group on Information Information Handling,[5] a coalition of partners addressing the information literacy of researchers and their supervisors, is also tackling the acquisition of data management skills. Library representative bodies are in this group, but the audience is not librarians. Domain knowledge acquired through a first degree in a science, technology or engineering field, is likely to be advantageous in working in data informatics and this has implications for the entry requirements and selection of students to LIS courses.

So are there barriers to recruiting to informatics? The European Bioinformatics Institute (EBI) employs many informaticians and there are professional groups like the International Society for Biocuration,[6] which promote the key role played by informatics experts. However recognition for these individuals is limited, and this has significant implications on career progression. If there is little credit, recognition or reward for informatics experts, then there is little incentive to follow this career path. All of this is in stark contrast to the view that there is a critical shortage of data informatics skills. Three potential 'jump-start' actions are proposed below:

1. **Deconstruct and define the core components** of data informatics and research data management for the curriculum of UK Library and Information Science.
2. **Analyse the entry qualifications and subject expertise** of students applying for LIS school places, with a view to increasing Science, Technology, Engineering, and Mathematics (STEM) entrants.

[4] International Curation Education (ICE) Forum: http://www.jisc.ac.uk/whatwedo/programmes/preservation/iceforum

[5] RIN Working Group on Information Handling: http://www.rin.ac.uk/our-work/researcher-development-and-skills/information-handling-training-researchers/working-group-i

[6] International Society for Biocuration: http://biocurator.org/

 doi:10.2218/ijdc.v7i1.220

3. **Set up an international Data Informatics Working Group** to explore how informatics methods, publications and expertise, can be promoted, recognised and rewarded as first-class outputs within the research community.

Library as a Five-Star Scholarly Communications Advocate

Scholarly communication is undergoing a slow but relentless revolution. The scholarly knowledge cycle described in 2003 (Lyon, 2003), which featured open access repositories linking research outputs with learning and teaching resources, is growing in credibility. There is now increasing momentum behind sharing research data, software and program code, gene sequences, models, simulations and methods, all driving towards 'reproducible research' (Peng, 2011). Recent workshops, such as Beyond the PDF,[7] and a Schloss DagStuhl workshop,[8] led to the creation of the Force 11 Community,[9] and have tried to crystallise ideas around publishing data-intensive research outputs, their description, discovery, citation and reuse, and to use light-weight Web-based tools to produce exemplars which demonstrate the capability of these new publishing platforms.

One approach proposed by David Shotton, University of Oxford, describes the Five Stars of Online Journal Articles (Shotton, 2011) as a benchmark for enhancing and enriching the scholarly publication process (see Figure 2 below).

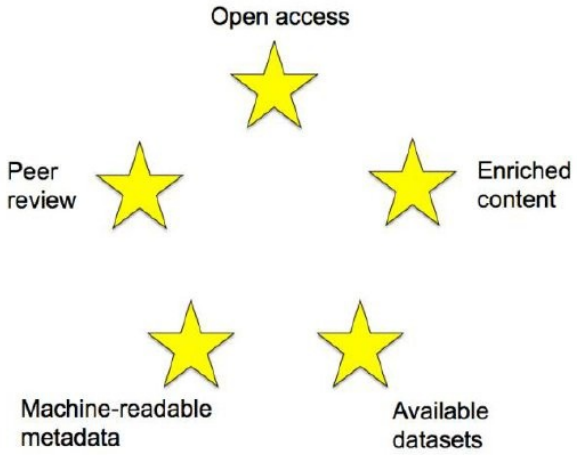

Figure 2. The Five Stars of Online Journal Articles.

[7] Beyond the PDF Workshop: http://sites.google.com/site/beyondthepdf/

[8] Future of Research Communication, Schloss Dagstuhl Perspectives Workshop: http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=11331

[9] Force 11 Community: http://www.force11.org/

The five stars in detail are:

- Peer review: Ensure your article is peer reviewed, to provide assurance of its scholarly value, quality and integrity.
- Open Access: Ensure others have cost-free open access both to read and to reuse your published article, to ensure its greatest possible readership and usefulness.
- Enriched content: Use the full potential of Web technologies and Web standards to provide interactivity and semantic enrichment to the content of your online article.
- Available datasets: Ensure that all the data supporting the results you report are published under an open licence, with sufficient metadata to support their re-interpretation and reuse.
- Machine-readable metadata: Publish machine-readable metadata describing both your article and your cited references, so that these descriptions can be discovered and reused automatically.

These five (deconstructed) points make highly compelling advocacy statements for the Library to promote and promulgate, to help embed more open scholarly communication practice amongst the academic community. Libraries are already active in promoting open access to publications through institutional repositories; the dissemination of strong advocacy statements of this sort together with potential new roles within libraries focusing primarily on scholarly communications, help to transform behaviours and facilitate cultural change. These elements are also components of the Community Capability Model (CCM) Framework,[10] which aims to capture the requirements for successful data-intensive research with a view to building capacity and capability within the sector. The CCM is being developed by UKOLN and Microsoft Research Connections.

Library as a Citizen Science Hub?

In 2009, the author spoke at a conference organised by the Local Government Association and the Museums, Libraries and Archives Council (MLA), which explored Modernising the Public Library Service. The presentation *Serving digital citizens: public libraries in the 21stC* (Lyon, 2009b) included a proposal that public libraries should become citizen science hubs, and act as an information and engagement point for the public to participate in citizen science activities. Since then, initiatives such as Galaxy Zoo and BBC LabUK have demonstrated real value in harnessing the effort, intelligence and contributions of the wider public. In addition, Research Councils UK (RCUK) have continued to demonstrate the importance of public engagement, through targeted funding calls (RCUK, 2011).

UKOLN worked with the British Library and the Association of Medical Research Charities on the JISC-funded Patients Participate! Project.[11] This examined the value of the lay summary, as a mechanism to bridge the gap between information access and

[10] Community Capability Model for Data-Intensive Research: http://communitymodel.sharepoint.com/Pages/default.aspx

[11] Patients Participate! Project: http://blogs.ukoln.ac.uk/patientsparticipate/

 doi:10.2218/ijdc.v7i1.220

public understanding of health-related information. The project explored the feasibility of creating crowd-sourced lay summaries for peer-reviewed articles in UK PubMed Central (UKPMC), and focused on stem cell research. The findings showed strong support for a lay summary to be published with every UKPMC article (PLoS Medicine already includes a lay summary with every article) and that guidelines and templates would be valuable in the creation of lay summaries. UKOLN and researchers from the Centre for Regenerative Medicine at the University of Bath, developed guidelines for researchers to assist the production of lay-summaries. This study suggests that libraries have an exciting opportunity to support public engagement with science in various ways:

- **Act as a hub** or one-stop shop and collate links and information about diverse citizen science activities
- **Support the researcher** in the production of lay summaries by advocating the use of guidelines and templates
- **Provide training** for new postgraduates through Doctoral Training Centres
- **Mediate public access** to research datasets in institutional, disciplinary or national repositories and data centres
- **Support members of the public** who wish to contribute and participate in science and research projects.

Looking Forward

This article is very much a personal perspective on the future of the library and focuses primarily on supporting research.

Research data informatics has been 'deconstructed' in different contexts including within the institutional research lifecycle, the delivery of new library services and understanding institutional roles and responsibilities. It is encouraging to note that there is pioneering work already underway in the US, Australia and the UK. Many practical examples, e.g. providing Web pages to support data management planning, are described in more depth, in an important new book (Corrall, 2012)

We should collectively view the 'Informatics Transform' as an important opportunity for libraries: a chance to engage new audiences, deliver innovative services and adopt new roles.

Author's Note

An earlier version of this article was submitted to the Royal Society Roundtable on the Future of Libraries in October 2011, which was a component of the major study Science as a Public Enterprise,[12] to examine the use of scientific information as it affects science, businesses and society. The project had a particular focus on the benefits from greater sharing of scientific information.

[12] Royal Society, Science as a Public Enterprise: http://royalsociety.org/policy/projects/science-public-enterprise/

Acknowledgements

UKOLN is funded by the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils as well as by project funding from the JISC, the European Union and Microsoft Research Connections. UKOLN also receives support from the University of Bath where it is based.

References

Borgman, C.L. (2011). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology.* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1869155

Corrall, S. (2012). Roles and responsibilities: Libraries, librarians and data. In G. Pryor (Ed.) *Managing Research Data (*pp. 105-133). London: Facet.

Hey, T., Tansley, S. & Tolle, K. (Eds). (2009). *The fourth paradigm: Data-intensive scientific discovery.* Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Hey, T. & Trefethen, A. (2003). *The data deluge: An eScience perspective.* Retrieved from http://eprints.ecs.soton.ac.uk/7648/1/The_Data_Deluge.pdf

Lyon, L. (2009). *Open science at web-scale: Optimising participation and predictive potential.* Consultative Report. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#november-2009

Lyon, L. (2009b). Serving digital citizens: Public libraries in the 21stC? Paper presented at the LGA/MLA Conference, London, UK. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html#2009-12-14-LGA-london

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities and relationships.* Consultancy Report. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#2007-06-19

Lyon, L. (2003). eBank UK: Building the links between research data, scholarly communication and learning. *Ariadne, 36.* Retrieved from http://www.ariadne.ac.uk/issue36/lyon/

Nature. (2008). Nature: Big data issue 4. Retrieved from http://www.nature.com/news/specials/bigdata/index.html

Pain, E. (2011). Science careers: Surfing the tsunami. *Science.* http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2011_02_11/caredit.a1100013

 doi:10.2218/ijdc.v7i1.220

Peng, R.D. (2011). Reproducible research in computational science. *Science, 334.* Retrieved from http://www.sciencemag.org/content/334/6060/1226.full

Pennisi, E. (2011). Will computers crash genomics? *Science, 331*(6018). Retrieved from http://www.sciencemag.org/content/331/6018/666.short

Shotton, D. (2011). *Five stars of online journal articles.* [Blog post]. Retrieved from http://opencitations.wordpress.com/2011/10/17/the-five-stars-of-online-journal-articles-3/

Research Councils UK. (2011). *RCUK Public Engagement with Research: CATALYSTS call for proposals.* Retrieved from http://www.rcuk.ac.uk/per/pages/catalysts.aspx