IJDC | *General Article*

Collection, Curation, Citation at Source: Publication@Source 10 Years On

Jeremy G. Frey
University of Southampton

Simon J. Coles
University of Southampton

Colin L. Bird
University of Southampton

Cerys Willoughby
University of Southampton

Abstract

The Southampton chemical information group had its genesis in 2001, when we began an e-Science pilot project to investigate structure-property mapping, combinatorial chemistry, and the Grid. CombeChem instigated a range of activities that have since been underway for more than ten years, in many ways matching the expansion of interest in using the Web as a vehicle for collection, curation, dissemination, reuse, and exploitation of scientific data and information. Chemistry has frequently provided the exemplar case studies, notably for the series of projects – funded by Jisc and EPSRC – that investigated the issues associated with the long-term preservation of data to support the scholarly knowledge cycle, such as the eBank UK project.

Rapid developments in Internet access and mobile technology have significantly influenced the way researchers view connectivity, data standards, and the increasing importance and power of semantics and the Semantic Web. These technical advances interact strongly with the social dimension and have led to a reconsideration of the responsibilities of researchers for the quality of their research and for satisfying the requirements of modern stakeholders. Such obligations have given rise to discussions about Open Access and Open Data, creating a range of alternatives that are now technically feasible but need to be socially acceptable. Business plans are changing too, but in a strange contradiction, desire can run ahead of what is possible, sensible, and affordable, while lagging behind in imagination of what would be technically possible and potentially game-changing!

Taking the chemical sciences as our example and focusing on the curation of research data, we explore from our perspective, ten years back and ten years forward, how far we have been able to re-imagine the data/information value pathway from bench to publication. We assess not only the major advances and changes that have been achieved, but also where we have been less successful than we might have hoped. We explore the directions for the future, based on what is clearly already possible and on what we can envisage becoming feasible in the near future.

Received 12 December 2014 | *Accepted* 10 February 2015

Correspondence should be addressed to Professor Jeremy Frey, Chemistry, University of Southampton, Southampton SO17 1BJ. Email: j.g.frey@soton.ac.uk

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

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. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The Southampton chemical information group had its genesis in 2001, when we began an e-Science pilot project to investigate structure-property mapping, combinatorial chemistry, and the Grid. CombeChem (Frey et al., 2006) instigated a range of activities that have since been underway for more than ten years, in many ways matching the expansion of interest in using the Web as a vehicle for collection, curation, dissemination, reuse, and exploitation of scientific data and information. Chemistry has frequently provided the exemplar case studies.

Describing the activities of the Jisc-funded eBank UK project[1], Lyon used the CombeChem project and the EPSRC National Crystallography Service (NCS)[2] as examples to illustrate her depiction of the scholarly knowledge cycle (2003). Her premise was that the accumulation of knowledge is based on the continuing use and reuse of data and information, such that research and learning entail cyclical processes. Although her depiction of the cycle does not refer specifically to collection, curation, or citation, it is apparent that all three processes are intrinsic to the workflows. For the cycle to operate effectively and efficiently, it is good laboratory practice for data creators to anticipate what researchers might encounter in later stages by collecting, curating, and preparing for citation as they create the data, at source (Frey, 2008). Figure 1 illustrates the processes in the context of the research and teaching lifecycle.

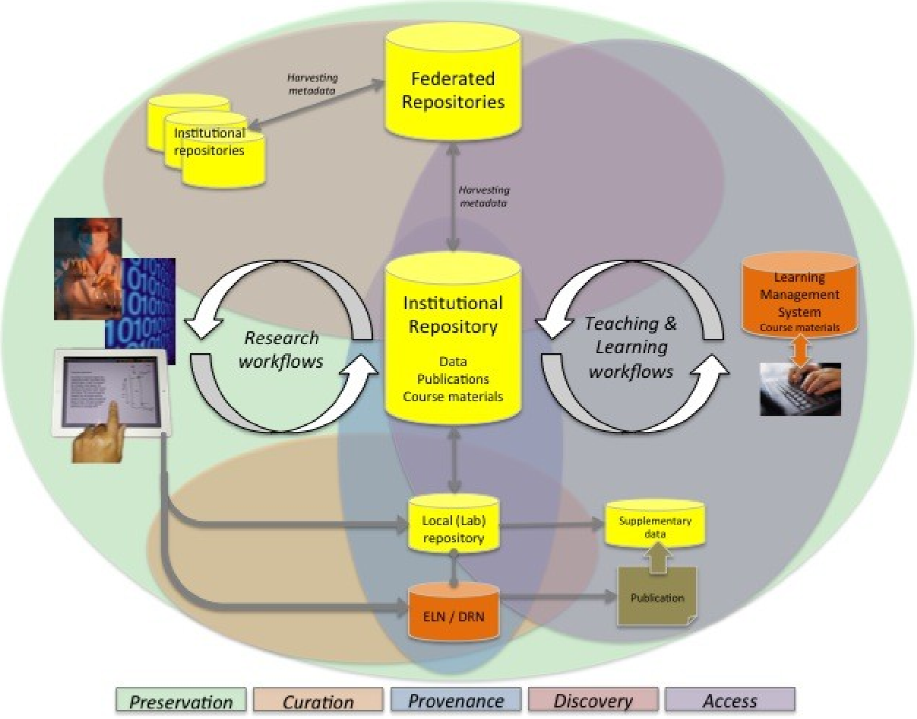

Figure 1. Aspects of the curation lifecycle for research and teaching. Adapted from the Scholarly Knowledge Cycle (Lyon, 2003).

[1] eBank UK: http://www.jisc.ac.uk/whatwedo/programmes/eresearch/semanticgrid/ebankuk.aspx

[2] National Crystallography Service: http://www.ncs.ac.uk/

In 2002, Frey et al. introduced the concept of publication at source (2002), aiming to highlight the responsibilities of researchers for the dissemination of their data and results in addition to contextual information about the lifecycle of the experiment. They argued that the developing Web would give to those researchers the potential to ensure that the products of their research could be described correctly in context. That context should be supplied by those best placed to provide high quality metadata, with the assurance that the context would be maintained and augmented by transforming data and information to knowledge and wisdom.

Contextual lifecycle information is precious not only for reproducibility, which is a fundamental test of research integrity, but also when comparing and contrasting the results of other experiments. On reflection, however, we are obliged to regard the propagation of context as an area where neither the community nor we have achieved what we aspired. Scientific publications, conventional and online, still omit contextual information. Few systems capture context in a form that is machine-processable, using, for example, RDFa mark-up[3].

The ten years that followed saw rapid developments in Internet access and mobile technology, which have significantly influenced the way researchers view connectivity, data standards, and the increasing importance and power of semantics and the Semantic Web. These technical advances interact strongly with the social dimension and have led to a reconsideration of the responsibilities of researchers for the quality of their research and for satisfying the requirements of modern stakeholders (The Royal Society, 2012).

Our empirical indications are that physical sciences students begin their courses with a poor understanding of the scientific method and of good research practice; often they are lacking critical analysis skills. There is evidence that such concerns are also present at doctoral level. The 2012 report from Jisc and the British Library revealed that relatively few doctoral science students were using large datasets as primary sources and students in other disciplines preferred text-based and secondary sources (2012). The report also asserts:

> ‘Training for research work and for information use is an area of overall dissatisfaction among Generation Y doctoral students.’

In our recent feature article for Information Standards Quarterly (Bird et al., 2013a), we discuss data curation issues in the chemical sciences and examine data curation in practice. While we assert that curation should be a fundamental aspect of research processes, we also acknowledge the need for tools that encourage and facilitate the recording of context at source in the form of appropriate metadata. We have developed LabTrove, a researcher-centric Electronic Laboratory Notebook (ELN) and recently reported the experiences of researcher using LabTrove in a heterogeneous set of academic laboratories (Badiola et al., 2015). Given the premise that the accumulation of knowledge depends upon the reuse of data and information, protocols for discovery, access, and processing are necessary, to which end we have proposed the elnItemManifest, a metadata schema for describing at a high level the knowledge held in ELNs (Coles et al., 2013).

There can be little argument that the currency of collection, curation, and citation is metadata. In the next section we examine the role of metadata in the knowledge cycle and assess the implications of our investigations into how user-defined metadata can create context for the experiment record (Willoughby, 2014). Given our emphasis on

[3] RDFa Core 1.1 – Third Edition: http://www.w3.org/TR/rdfa-syntax/

recording context at source, we examine the question of who is ultimately responsible for curation, and appraise where gaps can open in the lifecycle of collection, curation, and citation. We consider the consequences of deferring care for data.

Although much has been written about *approaches* to curation, notably by the Digital Curation Centre (2014) and the Research Information Network (2008), it would seem that *attitudes* to curation have received much less attention. Investigators involved with the Human Brain Project published a perspective on data sharing (Gardner, 2003), in which they advise careful but selective curation (although they do not use that term as such):

> ‘Policies should recognize that small amounts of adequately characterized, focused data are preferable to large amounts of inadequately defined and controlled data stored in a random repository.’

Coles et al. make a similar point, noting that journal publications tend to rely on the most pertinent results, usually in a summary form only. This uncoupling of the publication from the experimental data “renders replication or reuse of the data impossible and results in severe information loss” (2007).

Other issues endure over which research groups and individual researchers have little, if any, control. Raw data files that in principle should be attached to a publication can be uncomfortably large, spectra being a prime example. Moreover, upload restrictions can oblige researchers to rely on other, discrete, repositories. Regrettably, the design of some such data repositories allows only for higher-level abstractions, for example, the crystal structure but not the raw X-ray data. Some instrument manufacturers continue to rely on proprietary formats for data files, which can negate the advantage of storing and sharing these files.

Having assessed the major advances and changes since the genesis of CombeChem, we explore the directions for the future, based on what is clearly already possible and on what we can envisage becoming feasible in the near future. One early finding was that a directive to “make your data available to others”, for example on a site within the researcher’s control, has a very different appeal to researchers and creates a much more favourable response compared to a request to “give your data to the publisher or even to the library”. Some attitudes have eased a little. Previously, researchers wanted to keep all their data safe, by which they meant on a hard disk under their desk or, if possible, as a printout. The increasing size of datasets and the growing risk of loss as the result of the catastrophic failure of a disk (or a fire as an extreme example), has led to researchers becoming happier with cloud storage, perhaps too uncritically so. Cloud storage does at least facilitate future access to the data, but does nothing directly for the quality that will enable intelligent access. The move by libraries and data stores to obtain but embargo data allows them to believe that eventually they will be able or allowed to integrate those resources. Time will tell if this is in fact a realisable vision or an illusion created to satisfy the grant authorities.

The Role of Metadata

While definitions of the term can, and do, depend on one’s perspective, metadata is indispensable to the lifecycle of data, information, and knowledge. It is essential for effective sharing, reuse, and dissemination. In anticipation of the eventual reuse and citation of data and information, the metadata should be collected and the context curated at source. Re-deriving context at some later stage is commonly a complex, expensive, and error-prone process (Frey, 2008). Michener et al. (1997) captured the fundamental justification nearly 20 years ago:

> ‘The most important reason to invest time and energy in developing metadata is that human memory is short.’

We might add that human memory is frequently inaccurate and the human mind has a tendency to interpolate information with more regard to making a story than reproducing the exact sequence of events, as typified by problems with eyewitness testimony in court trials (Engelhardt, 1999):

> ‘… the mere fault of being human results in distorted memory and inaccurate testimony.’

In our appraisal of data curation issues in the chemical sciences, we set out the nature and capabilities of metadata, prior to expressing our view that capturing context is its single most important function (Bird et al., 2013a):

> ‘Problems can and do arise later in the research cycle if researchers do not capture the correct context as they record their experiments and acquire their data. When reviewing a research project for any purpose, such as analysis, publication, or to reproduce the results, it is crucial to be able to appreciate the full context of the data and information.’

Frey has examined the potential of Semantic Web technologies in managing and exploiting data and information during the lifecycle of research in the chemical sciences (2009). Semantics improve selectivity when identifying relevant information; semantic metadata can be exploited in all three phases of a typical research project: planning, enactment, and dissemination, as illustrated not only by the Southampton projects but also the work of other groups described in the article. More recently, Frey and Bird have reviewed the contributions of the Semantic Web to the field of cheminformatics, which embraces all aspects of the management, sharing, and analysis of chemical data (2013). They highlight the importance of provenance and consider how the union of cheminformatics and the Semantic Web can enhance provenance. Both humans and computers can handle information that is captured in context with semantic metadata, acknowledging that the human researcher’s view can, and usually will, deviate significantly from the requirements of the computer system. This echoes the earlier thinking of Pancerella et al., who conclude their interpretation of metadata by stipulating that it must be in a machine-comprehensible format to enable it to be understood and manipulated (2003). This point of view has clear implications for curation: the metadata should be captured from the context at time of conducting the experiment, that is, at source. However, achieving this aim is still a major problem.

Form-filling systems that control the acquisition of metadata by requiring extensive tables to be completed either collect limited information as users try to avoid filling in the gaps or at best collect only the metadata that was foreseen as important. The outcome is that creativity is crushed and important information is left unrecorded. However, giving flexibility to users has the predictable effect that some authors will be creative and imaginative, while the essential but tedious context will remain as vague as ever. The argument that context information should be inferred automatically breaks down because automated capture is still beyond the capabilities of the typical software systems in use in chemistry laboratories.

Frey also argues that tools to capture and maintain the context of data will need to work well in the laboratory if the full capabilities of Semantic Web technologies are to be realised (2009). In this respect, ELNs and institutional repositories are fundamental to the collection, curation, and preservation of experiment data and information.

In 2006, Milsted et al. began to develop an ELN with the perspective of the individual researcher in mind, aiming to assist laboratory scientists to capture the planning and enactment of their research, appropriately marked up to facilitate its eventual dissemination – “Publication@Source”. Their system was web-based, with an underlying blog technology enhanced to provide features such as access control, recording templates, and flexible metadata support (2013). The ELN is now known as LabTrove and is used regularly in a number of institutions and for a broad range of research activities (Badiola, 2015). Although a notebook, whether digital or paper, is the customary medium for recording the research narrative, social media systems are in principle also good for the narrative but they are very poor for data management and registering provenance.

In line with the formal Oxford Dictionary definition of dissemination as “the act of spreading something, especially information, widely; circulation”[4], we take a very wide view of all aspects of the transfer of information and, with regard to current concerns about the role of government-funded research, of impact. The concepts of dissemination and impact are much broader than publication, engendering an obligation on researchers to share the fruits of their research to enable other scientists to make progress by building on those results. It is now expected that the results (the paper and the data) of publicly funded research will be made available in open repositories.

The EPSRC-funded Dial-a-Molecule Grand Challenge Network[5] realised that to achieve its aim of reducing significantly the time taken to develop new chemical compounds, it would be necessary to exploit the vast body of prior knowledge about reaction outcomes. The majority of the information required is in ELNs and currently is inaccessible. Even when available, it would still be necessary to develop protocols for discovery, access and ultimately automatic processing.

The Dial-a-Molecule network set up a working group to consider the issues relating to ELNs in effect “publishing” their content. The group put forward a three-layer model, comprising *knowledge*, *information*, and *processing* (or *data*) layers. Coles et al. set out the thinking behind this model, and describe the *elnItemManifest* schema that they propose for the *knowledge* layer (2013). The *elnItemManifest* consists of metadata that describes the content of an ELN record at a level that is succinct but sufficient to enable a prospective user to assess whether to request further detail and the contact information for doing so; in essence it “sets out the stall”.

Coles et al. also provide as demonstrators two *elnItemManifest* files generated in one case from LabTrove, in the other case the IDBS ELN. As the researcher captured the metadata provided when creating the ELN record, we can deem this automatic publication to be an example of relieving the *burden of curation*. Although this phrase used widely, its origin is unclear. However, we know from our own research that lack of understanding of the significance of curation, lack of awareness of terms and vocabularies, and a reluctance to disclose are all inhibitory factors that contribute to the perception that curation is a burden (Willoughby, 2014).

[4] OED - Dissemination: http://www.oxforddictionaries.com/definition/english/dissemination

[5] Dial-a-Molecule EPSRC Grand Challenge Network: http://www.dial-a-molecule.org/wp/

Our research investigated patterns of use and user attitudes towards metadata within LabTrove. We compared the usage in LabTrove with a variety of other platforms that support the creation of user-defined metadata. Our investigations revealed both similarities and differences between the platforms, enabling us to identify three approaches that could be adopted to encourage and support the creation of metadata by the community. Our findings have already influenced our ELN activities at Southampton.

Responsibilities: Who Should "Mind the Gap"?

Even a cursory examination of the scholarly knowledge cycle propounded by Lyon shows that duties for the collection, curation, or citation exist at every stage. The responsibilities of researchers for meeting the requirements of sound governance and ensuring the quality of their work have become more apparent and might arguably have increased with the expansion of Open Data and the promotion of Open Access.

Frey places the responsibility for curation firmly with the originator of the data, making the point that researchers who publish their results but fail to make the supporting information discoverable and thus reusable thereby lose an opportunity to expose their work (2008). He urges researchers to organise their data and preserve it with semantically rich metadata, captured at source, to provide short- and long-term advantages for sharing and collaboration. Frey's earlier view that it might be the responsibility of archivists to maintain the data, while the researcher retains responsibility for the information, has to some extent been overtaken by events (Frey et al., 2002). The Royal Society report asserts that researchers should accept responsibility for the quality of their research and for satisfying the requirements of modern stakeholders (2012) and the creators of data are now expected to preserve their results in open repositories. Such pressures place the responsibility on researchers themselves.

The commonly held view that computer systems exhibiting artificial intelligence (AI) will soon be powerful enough to extract the meaning out of any documents, thus rendering it unnecessary to be explicit about context, misses several key points. Even humans have trouble inferring the context in poorly written material and, while software has advanced, the main source of progress has been having large corpora of well marked-up materials with which to compare a given document. Once such a corpus exists and comparisons are possible – Google Translate[6] being just one example – tools to help with context acquisition could be particularly useful "at source", as the data is created, as well as for retrospective use.

Although the complexity and heterogeneity of chemical data can present some specific challenges, chemistry is by no means alone in addressing issues associated with collection, curation, and citation. To cite just one example, a case study of the Neuroimaging Group in the University of Edinburgh, as part of the SCARP project, found that data management facilities were needed at the laboratory level to mitigate the risks if information about context and provenance were not properly recorded. The study also identified a need for laboratory researchers to curate their data as well as share it, to enable other workers with differing skills or specialities to access and reuse the information (Whyte et al., 2008).

[6] Google Translate: https://translate.google.com/

> ‘As one would expect of an active research group, much curation and preservation activity is embedded in other research roles, particularly Principal Investigators who as custodians are responsible for clinical data management and security.’

Whyte et al. are clearly alert to the gaps that can open in the lifecycle of collection, curation, and citation, and have recognised the options for change that would enable researchers to “mind the gap”. The conspicuous opportunity for gaps to open occurs when those responsible for curation defer the activity until some later, ostensibly more convenient, time but then encounter the difficulties identified by Buneman et al. (2006):

> ‘Curators usually attempt to add links to the original publications or source databases, but in practice, provenance records are often absent, incomplete or ad hoc, often despite curators’ best efforts. Also, manually managed provenance records are at higher risk of human error or falsification.’

Given that metadata is the currency of collection, curation, and citation, the attitudes of chemists towards metadata are clearly significant for “minding the gap”. Our own research has highlighted the following issues (Willoughby, 2014):

- The lack of defined metadata schema,
- The lack of knowledge about metadata,
- The effort involved in creation,
- The lack of visibility and perceived benefits of metadata.

We recognise the pressing need for training and education to encourage researchers to curate the data as they collect it, that is, at source. The disturbing aspect of this recent evidence is that despite our efforts and those of other teams this century and before, metadata capture remains an issue for both the collection and the curation of data, so inevitably for citation too. Metadata issues are a clear example of where we have been less successful than we might have hoped. Although we have made progress with the automated collection of instrument metadata, the automatic capture of descriptive metadata remains a challenge.

At its worst, deferring curation can lead to the phenomenon of “lost knowledge”, exemplified by workers who leave or retire without imparting their expertise (DeLong, 2006). Moreover, as Michener et al. recognised, even the original creator of material is likely to forget the exact context over time, and have trouble finding the information they need at a later time (1997).

The phenomenon also manifests as “lost knowhow”, for example when a researcher has to redraw a graph without access to the full data or processing information used originally by a colleague who has since left the group.

Futurology?

Taking the chemical sciences as our lens, we have explored the curation of research data over the past 10 to 15 years from our perspective. During this period we have reassessed the value of data and information along the pathway from bench to publication. In assessing the major advances and changes that the wider community and we have achieved, we must acknowledge the growing recognition of the importance of quality metadata, especially semantically rich metadata. Tools that exploit the technologies of the Semantic Web are steadily becoming available, as are techniques for the automated capture of descriptive metadata, particularly from scientific instruments. Nevertheless, there will be a continuing need for more and better tools, as standards evolve rapidly and render many of the existing tools obsolete and unsupported.

However, it is in the capture at source of user-defined metadata that we have been less successful than we might have hoped. Electronic Laboratory Notebooks (ELNs) have and will continue to have an influential role in the research lifecycle and in the collection, curation, and citation of research outcomes (Bird et al., 2013b). We anticipate that ELNs will evolve into generic digital research notebooks, with markedly improved capabilities for curation and information sharing.

The issues we have outlined are as real today as they were ten years ago, but what we could not then foresee was the huge impact of the social media “explosion”, which has produced Twitter, LinkedIn, and Facebook, to give just three current examples; others have already faded from memory. The apparently free and easy availability of information about almost everything has stimulated the open access debate. In publication terms that dialogue is primarily about who pays for what, and publication business models in general; it leads to consideration of the crucial importance of access to data.

In considering directions for the future, we note that the chemical sciences span pure research, with no immediate financial return, through to the lucrative life sciences research and new materials. Both areas have very significant commercial potential and are susceptible to a wide range of human responses to socio-technical issues. Semantic support is extending and improving, automation is playing an even greater role in science laboratories, and new technologies are emerging, such as additive manufacturing (3D printing), which have the potential to transform research and scientific practice over the coming decade. Notwithstanding the prospects of what we can envisage becoming feasible in the near future, we must keep in mind the fundamental tenet of the research lifecycle (Bird et al., 2013a):

> ‘All science is strongly dependent on preserving, maintaining, and adding value to the research record, including the data, both raw and derived, generated during the scientific process. This statement leads naturally to the assertion that all science is strongly dependent on curation.’

Acknowledgements

This work was funded by the RCUK e-Science programme (EPSRC grant GR/R67729, EP/C008863, EP/E502997, EP/G026238, BBSRC BB/D00652X), the EPSRC National Crystallography Service, the HEFCE and JISC Data Management Programme and the University Modernisation (UMF), and the RCUK Digital Economy Theme, IT as a Utility Network+ (EPSRC EP/K003569).

References

Badiola, K., et al. (2015). Experiences with a researcher-centric ELN. *Chemical Science, 6,* 1614–1629. doi:10.1039/C4SC02128B

Bird, C., Willloughby, C., Coles, S., & Frey, J. (2013a). Data curation issues in the chemical sciences. *Information Standards Quarterly, 25*(3), 4–12. doi:10.3789/isqv25no3.2013.02

Bird C., Willoughby C., & Frey J. (2013b). Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences. *Chemical Society Reviews, 42,* 8157–8175. doi:10.1039/C3CS60122F

Buneman, P., Chapman, A., Cheney, J., & Vansummeren, S. (2006). A provenance model for manually curated data. In L. Moreau and I. Foster (Eds.), *Lecture Notes in Computer Science: Vol. 4145. Provenance and Annotation of Data* (pp. 162–170). doi:10.1007/11890850_17

Coles, S., Carr, L., & Frey, J. (2007). *R4L: The repository for the laboratory.* Retrieved from http://r4l.eprints.org/publications/docs/R4Lfinalreport.pdf

Coles, S., Frey, J., Bird, C., Whitby, R., & Day, A. (2013). First steps towards semantic descriptions of electronic laboratory notebook records. *Journal of Cheminformatics, 5,* Article 52. doi:10.1186/1758-2946-5-52

DeLong, D.W. (2006). *Lost knowledge: Confronting the threat of an ageing workforce.* New York, NY: Oxford University Press.

Digital Curation Centre. (2014). *What is digital curation?* Retrieved from http://www.dcc.ac.uk/digital-curation/what-digital-curation

Engelhardt, L. (1999). The problem with eyewitness testimony: A talk by Barbara Tversky, Professor of Psychology, and George Fisher, Professor of Law. *Stanford Journal of Legal Studies, 1*(1), 25–29. Retrieved from http://agora.stanford.edu/sjls/Issue%20One/problem.htm

Frey, J. (2008). Curation of laboratory experimental data as part of the overall data lifecycle. *International Journal of Digital Curation, 3*(1), 44–62. doi:10.2218/ijdc.v3i1.41

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

Frey, J., & Bird ,C. (2013). Cheminformatics and the Semantic Web: Adding value with linked data and enhanced provenance. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 3,* pp. 465-481. doi:10.1002/wcms.1127

Frey, J., De Roure, D., & Carr, L. (2002). *Publication at source: Scientific communication from a publication Web to a data Grid.* Paper presented at the Euroweb 2002 Conference, Oxford, UK. Retrieved from http://ewic.bcs.org/content/ConWebDoc/4084

Frey, J., De Roure, D., Taylor, K., Essex, J., Mills, H., & Zaluska, E. (2006). CombeChem: A case study in provenance and annotation using the Semantic Web. In L. Moreau and I. Foster (Eds.), *Lecture Notes in Computer Science: Vol. 4145. Provenance and Annotation of Data* (pp. 270–277). doi:10.1007/11890850_27

Gardner, D., Toga, A. W., Ascoli, G. A., Beatty, J. T., Brinkley, J. F., Dale, A. M., … Wong, S. T. C. (2003). Towards effective and rewarding data sharing. *Neuroinformatics, 1,* 289–295. doi:10.1385/NI:1:3:289

Jisc & British Library. (2012). *Researchers of tomorrow: The research behaviour of Generation Y doctoral students.* Retrieved from http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow.aspx

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/

Michener, W., Brunt, J., Helly, J., Kirchner, T., & Stafford, S. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications, 7,* 330–342. doi:10.2307/2269427

Milsted, A., Hale, J., Frey, J., & Neylon, C. (2013). LabTrove: A lightweight, web based, laboratory "blog" as a route towards a marked up record of work in a bioscience research laboratory. *PLoS ONE*, *8*(7), e67460. doi:10.1371/journal.pone.0067460

Pancerella, C., Hewson, J., Koegler, W., Leahy, D., Lee, M., Rahn, L., … Frenklach, M. (2003). Metadata in the collaboratory for multi-scale chemical science. In *Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications* (pp. 121–129). Dublin Core Metadata Initiative. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/740

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

The Royal Society. (2012). *Science as an open enterprise* (Science Policy Centre Report No. 02/12). Retrieved from https://royalsociety.org/policy/projects/science-public-enterprise/Report/

Whyte, A., Job, D., Giles, S., & Lawrie, S. (2008). Meeting curation challenges in a neuroimaging group. *International Journal of Digital Curation, 3*(1)*,* 171–181. doi:10.2218/ijdc.v3i1.53

Willoughby, C., Bird, C., Coles, S., & Frey, J. (2014). Creating context for the experiment record. User-defined metadata: Investigations into metadata usage in the LabTrove ELN. *Journal of Chemical Information and Modelling, 54*, 3268–3283. doi:10.1021/ci500469f