INFORMATION STANDARDS QUARTERLY
FALL 2013 | VOL 25 | ISSUE 3 | ISSN 1041-0031

ISQ

TOPIC

DATA CURATION

DATA CURATION ISSUES IN THE CHEMICAL SCIENCES

DATA CURATION IN THE OPENAIRE SCHOLARLY COMMUNICATION INFRASTRUCTURE

PRESERVING THE GREY LITERATURE EXPLOSION: PDF/A AND THE DIGITAL ARCHIVE

ENSURING THE LONG TERM IMPACT OF EARTH SCIENCE DATA THROUGH DATA CURATION AND PRESERVATION

NISO
How the information world CONNECTS

COLIN L. BIRD, CERYS WILLOUGHBY,
SIMON J. COLES, & JEREMY G. FREY

DATA CURATION

ISSUES IN THE CHEMICAL SCIENCES

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.[1]

Chemistry is no exception, and given the significance of chemical data to many other disciplines, we assert that curation should be a fundamental aspect of the research practice in the chemical sciences. In this article we investigate the extent to which chemists do actually respect the importance of curation in their day-to-day activities in the laboratory or, nowadays, frequently at the computer. »

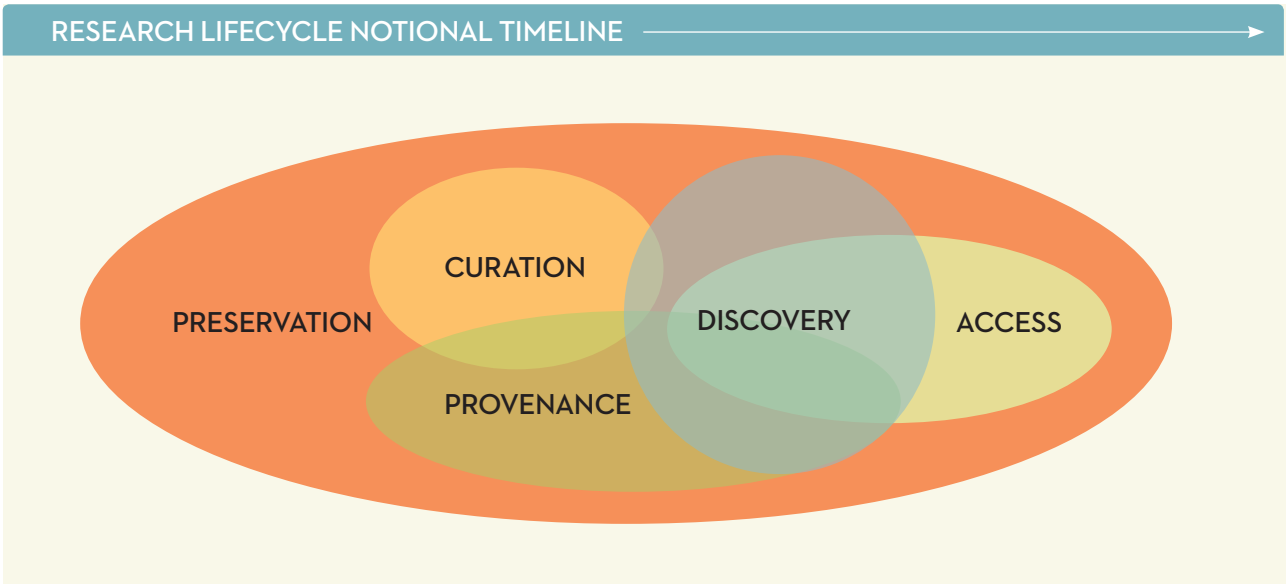

Figure 1: The concepts of preservation, curation, provenance, discovery, and access in the context of the research lifecycle *Reused with permission*[3]

Several of us have examined the origins and evolution of scientific record keeping in our recent review of laboratory notebooks in the digital era.[2] We compare the electronic[a] laboratory notebook (ELN) with the traditional paper notebook, noting that paper still has some advantages for maintaining a "journal" of research activities. However, digital recording is manifestly superior for capturing and annotating data, providing for machine and human readability, mobility, sharing, and greatly reducing the opportunities for error. Digital (electronic) recording also facilitates curation as part of the research lifecycle. Figure 1 illustrates how the concepts of preservation, curation, provenance, discovery, and access are embedded within the research lifecycle.

These five concepts are intimately related but chemists tend to consider them, if at all, separately and then only long after the initial actual capture of the data or information. It has always been the case that it is not sufficient to capture and preserve the outputs of chemistry research: curation is essential if those outputs are to be discoverable, accessible, and subsequently reusable. Governmental, professional, and funding agencies are understandably concerned to maximize the exposure and reuse of data and information, and thus the impact of public funding. Accordingly, these agencies are reminding researchers that preservation and curation are among their professional responsibilities when undertaking scientific research.

An essential ingredient in the curation process is metadata: descriptive information and classification labels that group related items, provide context, and facilitate the reuse of specified research outputs.

Zeng and Qin[4] present examples taken from a math dictionary and an educational library system to "demonstrate that metadata is capable of performing the following tasks:

- Describing what resources are and what they are about, and organizing those resources according to controllable criteria
- Allowing resources to be found by relevant criteria, aggregating similar resources, and providing pathways to the location of the desired information
- Facilitating metadata exchange and enabling interoperability
- Providing digital identification and description for archiving and the preservation of resources"

In our view, the single most important function of metadata is to capture context. 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.

In this article we review scientific curation practices, particularly those of chemists, to establish the extent to which those practices conform to the views of Zeng and Qin. We then consider the potential roles for librarians and information specialists in assisting with scientific curation, both directly and in training scientists—especially

[a] We note that the commonly used term "electronic" really refers to the recording of the research notes in a digital medium rather than the electronic nature of the devices often used to enter the information to the system.

chemists—to curate their outputs. The goals for chemists and other scientists will be to comprehend the importance of context and to capture appropriate and sufficient metadata in a timely and proportionate manner.

We also examine the key issue of the "burden of curation" and identify a set of curation challenges that will become increasingly important for chemistry, the related chemical sciences, and indeed other disciplines in the next decade. During this period, public expectations will grow with regard to transparency and the impact of research, while funding will remain level at best.

Chemistry faces some unique challenges, which will have a bearing on curation in practice. Data is typically complex and heterogeneous, reflecting an environment that is a mixture of the physical and the digital. Although outputs and observations are increasingly "born digital," many are first recorded in a non-digital form. Moreover, the potential for commercialization can lead to intellectual property issues; and chemists have an established conservative culture.

Curation in practice

If a paper notebook is complete and appropriately maintained, it is arguably self-curating, with the caveats that the nature and scope of the notebook contents are discoverable only if made public and are accessible only by inspection. The notebooks of eminent scientists such as Faraday[5] and Darwin[6] were notable for the thoroughness of their recording. Their paper notebooks contain thoughts, plans, procedures, observations, data, and calculations. The linking is intimate, as entries will be on the same page or associated by page reference; indexing facilitates discovery.

The use of digital formats for data—typically obtained from equipment, so intrinsically digital—and reports created a compelling incentive for the transition from paper to electronic recording, which we examined in our review of laboratory notebooks in the digital era.[2] We noted also that powerful computing facilities have become a necessity to keep pace with the expanding volume of data (the so-called data deluge[7]) and to retain control of results and other information: effective digital curation is now essential.

In 2008, Downing et al. conducted a survey of research chemists at Cambridge and Imperial College, intended primarily to investigate the factors influencing open data sharing. Significantly, they found a tendency to store data as hard copy.[8] It is a matter of observation, in laboratories across the world, that chemists continue to prefer paper notebooks and to store data on disparate systems linked to instruments or on their personal computers. Chemists frequently use proprietary software for data analysis and report writing, then preserve the resulting files on a variety of computing platforms and systems. Such files are manifestly difficult to discover and access; collaborative activities commonly depend on peer-to-peer communication by e-mail. Collaborators might be using different software, and often share derived information—possibly "cleaned up"—without also making the primary data available.

It is self-evident that such practices are no longer appropriate in the digital era, especially as funding bodies now mandate that all grant proposals include a data management plan, which must include provisions for data access by other researchers.[9] It is also self-evident that data management does and will require data and information curation. The level of curation required will necessarily go beyond description and classification. Semantic metadata will be necessary to enable machines to handle the relationships between data with differing characteristics. Notebooks, reports, and publications are essentially unstructured, whereas the data generated by instruments and as the result of computations is structured. That structure is known, but must be communicated as part of the curation process. We believe it to be fundamental to the future of science that semantic data integration becomes an innate part of curation. Two of us have recently reviewed the progress made by cheminformatics towards the goals of data integration, owing to the influence of Semantic Web technologies.[10]

Reproducibility is a basic tenet of scientific methodology, obliging researchers to provide sufficient information to enable verification of their work. Similarly, data management plans are but one instance of regulatory requirements that can give rise to audits. The so-called "Duke University scandal" strongly demonstrates the potential consequences of failing to provide adequate information for both audit and reproducibility. Ince attributes the affair in part to "sloppiness in data curation and software storage."[11]

While it is beyond the scope of this paper to give full coverage to the movement to support a much greater degree of reproducibility in scientific publications, we point the reader to the Reproducibility Initiative,[12] which follows on from the work on reproducible documents and papers[13] that began to attract attention in at the beginning of the decade, driven by researchers at Stanford University.

Du and Kofman took the view that improved data annotation was essential for ELNs "to minimize the possibility of misinterpretation, and encourage communication between users when the meaning of data needs to be clarified."[14] Although they do not use the word curation, it is clear that effective curation is what they are calling for. Ten years

CONTINUED »

Arguably, the two most apposite issues from the perspective of curation are the ever increasing breadth of knowledge required to deal with interdisciplinary research, even within problems specific to the chemical sciences, and the global scale of collaborative efforts.

earlier a similar point had been made by Michener et al., who contributed a perceptive justification for the curation of scientific data:[15]

> ***The most important reason to invest time and energy in developing metadata is that human memory is short. If data are to undergo any secondary usage, then adequate metadata will be required even if that secondary usage consists of reuse by the data originator.***

In this paper, we do not attempt to exhaustively define metadata, although we do concur with Pancerella et al. that the description data about data "is very dependent on one's perspective."[16] They conclude their own definition with a stipulation that has clear implications for curation: "…because metadata must be understood and manipulated, it must be formatted in a way that exposes its meaning in machine-comprehensible form." It seems inevitable that curation practice will increasingly emphasize the capture of metadata at the time data and information are created, which Frey describes as curation at source.[17]

We have to ask who will do the curating, particularly at source. It is unrealistic to expect laboratory chemists to adapt their working practices overnight; they will need education, training, and above all encouragement. Chemists will need to see some of the benefits immediately. Curation performed later in the research lifecycle might become or remain a task for professionals. For example, the Chemical Abstracts Service databases are curated and quality-controlled by scientists engaged by the service.[18] Nevertheless, Frey asserts that the best context and provenance is captured by the originators of the data:[17]

> ***Curation should be a matter of concern to the laboratory researcher. It should not be regarded as someone else's responsibility, nor undertaken at some late stage in the production of quality scientific data.***
>
> ***Gathering metadata when they are available is much simpler and cheaper than trying to remember or reconstruct them later.***

A role for libraries and information specialists

Losoff sets the scene for future transformations in electronic data management in her discussion of the role for librarians in data curation:[19]

> ***Scientific progress increasingly relies on searchable and intelligent integration of data sets, mined in conjunction with journals and other resources.***

Prior to the digital era, librarians would have seemed remote from the majority of scientists. The role of the librarian as curator would have been to manage book collections, probably using the Dewey classification system. Some librarians might have found a role assisting with indexing the contents of publications.

The developing role for librarians in the digital era is to provide a range of services associated with the management of data within their institutions. Swan and Brown report increasing calls on librarians to advise researchers and to provide practical help with data management.[20] They suggest that libraries can adopt a data care role:

> ***Many librarians are repositioning their libraries to take on the role of caring for data on behalf of the institution and data scientists we interviewed believe that libraries should indeed be responsible for data archiving and preservation. They believe this would free their own time to focus on working with researchers on different (domain-specific) data challenges.***

However, we believe that a sharp division of responsibilities is unlikely to achieve the most effective curation, especially in the chemical sciences, owing to the importance of curation at source. Chemists can learn valuable skills from librarians, but will need to engage with their libraries at a different level. Swan and Brown suggest that one of the main potential roles for libraries is to increase data awareness among researchers:[20]

> ***Libraries usually offer information literacy programmes to undergraduates, for example, but it is uncommon to see these penetrating the research programme and, if they do, they are very rarely compulsory. Nonetheless, it is likely that the data deluge will change things.***

Conversely, in our experience, the library community is apprehensive about the implications of this transition. Few librarians regard themselves as competent in data management, metadata, and practical curation. Advanced training in the associated skills is clearly necessary.[21] We foresee a valuable opportunity for librarians to work more closely with chemists, assisting with preservation, curation, and maintenance of provenance chains, even to the extent of "going native" and working with researchers in their laboratories.

Key issues associated with the burden of curation

Although the origin of the phrase *burden of curation* is unclear, its use is widespread.

From our own research into the behaviors of scientific communities and their patterns of metadata usage in ELNs,[22] we know that the following factors exacerbate the community perceptions of curation:

- Lack of understanding of the purpose and value of curation
- Reluctance to provide any more information than is deemed necessary
- Lack of consensus on the terms and vocabularies to use

Although Borgman does not refer to the *burden of curation*, she expresses the concept very clearly in her article entitled *Data, disciplines, and scholarly publishing*:[23]

> ***Second, and closely related, is the effort required to document data. Describing and tracking data for one's own use, and the use of labmates and other current collaborators, is far simpler than documenting them for use by unknown others. Even making data available for private exchange with known users requires richer explanations of the methods by which the data were collected, cleaned, analyzed, recorded, and interpreted. To make data available for public repositories, they may also need to be organized in compliance with community standards for metadata and ontologies.***
>
> ***Data curation is expensive for reasons similar to those which apply to publishing: peer review, editorial processes, technical support, and maintenance. Data also are much less 'self-documenting' than are publications. Without metadata and descriptions of research methods and of the context for data collection, they may simply be tables of numbers, lists of codes, pretty pictures, or boxes of rocks.***

In 2008, Jisc[b] commissioned a report into the costs and benefits to higher education institutions in the UK of preserving research data.[24] Acknowledging the burden of curation, the report called for investment in training people for data curation and preservation work, particularly metadata capture at the time of data acquisition and ingestion.

Given the importance of metadata in the curation process, our own research showed that the attitudes of chemists towards metadata highlighted several issues:

- 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

Challenges for curation in the chemical sciences

Recently, two of us reviewed information and data sharing in the chemical sciences.[3] In that review, we examined the progress made within the chemical sciences towards meeting the six changes set out in the recent Royal Society report as necessary for the pursuit of science as an open enterprise.[25] Concluding that review, we presented five areas that would present challenges in the context of defining and exchanging chemical information.

Arguably, the two most apposite issues from the perspective of curation are the ever increasing breadth of knowledge required to deal with interdisciplinary research, even within problems specific to the chemical sciences, and the global scale of collaborative efforts. The consequential challenge is to curate data and information to enable researchers from other disciplines to focus on why the material is relevant and important.

We referred earlier in this paper to the importance of reproducibility in science. To achieve reproducible observations, experiments, and computations, curation at source is more efficient and more effective than retrospective curation, especially with regard to reducing the risk of error.

We are in the era of "big data"; we have an increasing ability to create large amounts of data, in many cases as the result of automated processes. Hey and Trefethen foresaw in 2003 the implications of the "data deluge":[7]

> ***We therefore need to automate the discovery process—from data to information to knowledge—as far as possible. At the lowest level, this requires automation of data management with the storage and organization of digital entities. At the next level we need to move towards automatic information management. This will require automatic annotation of scientific data with metadata that describes both interesting features of the data and of the***

CONTINUED »

[b] Formerly the Joint Information Systems Committee (JISC), now a company promoting the use of digital technologies in UK education and research. See: www.jisc.ac.uk/about

> *storage and organization of the resulting information. Finally, we need to attempt to progress beyond structure information towards automated knowledge management of our scientific data. This will include the expression of relationships between information tags as well as information about the storage and organization of such relationships.*

To mitigate the burden of curation when managing "big data," automated methods will assume increasing significance. In our review, we assess the use of text- and data-mining techniques as tools for automated curation.[3] For the chemical sciences, such techniques will almost inevitably rely on structure-reaction-property databases, such as ChemSpider.[26]

As ever more data becomes available, expectations grow that science can solve problems faster. To meet such demands, the challenge for curation will be to facilitate faster locating of relevant material, for which the third change called for by the Royal Society is particularly pertinent: "the development of common standards for communicating data."[25]

Common vocabularies

Achieving general agreement over standards, particularly metadata vocabularies, is arguably the greatest challenge for all disciplines, not only for the chemical sciences. Reflecting on the problems associated with the long-term preservation of digital information, Hay and Trefethen emphasized the common cause:[7]

> *Needless to say, a solution to these problems is much more than just a technical challenge: all parts of the community from digital librarians and scientists to computer scientists and IT companies need to be involved.*

As set out by Zeng and Qin,[4] metadata is valuable for understanding data as well as for locating it. From our own research into patterns of metadata usage, it is apparent that metadata vocabularies and schema need to consider not only the wider research disciplines but also of individual research groups.[22] Flexibility is at least as important as control, if not more so.

Milsted et al. describe the development of a lightweight, researcher-centric ELN, for which they stress the importance of an extensible and flexible metadata framework:[27]

> *Our experience is that freedom, and the ability arbitrarily to add or modify metadata keys and values, is critical to the recording of what is planned and what actually occurs in the small scale experimental laboratory. In this sense, we distinguish between recording a procedure, and describing an experiment.*

Nevertheless, they do not reject the use of formal schemes:

> *In many such cases it will be appropriate to prompt the user to use terms from existing controlled vocabularies by providing templates that use those terms.*

> *Consistent usage creates the opportunity to map the informal vocabulary that arises out of local practice onto externally constructed ontologies and controlled vocabularies.*

As part of our review of information and data sharing in the chemical sciences, we examine metadata formats and chemical ontologies, with particular reference to the emergence of the Chemical Semantic Web.[3] We compare the attributes of the principal forms of controlled vocabulary, as illustrated schematically in Figure 2, reproduced from our review. We are currently planning a detailed review of the development of chemical ontologies and their applications in the chemical sciences.

Frey, discussing the value of the Semantic Web in the laboratory, notes that the "use of a controlled vocabulary ensures that everyone uses the same terms, but these terms have to be agreed and workable." However, he also points out that ontology construction can involve considerable effort.[28]

The development of ontologies for components of the chemical information space is clearly a necessary step. The Royal Society of Chemistry offers free downloads of subject classifications covering selected areas of chemistry.[29] Clearly, there are potential roles for other learned societies and for academic publishers in fostering agreement over vocabularies and schema for the chemical sciences. However, a further degree of common agreement is necessary to enable individual ontologies to be used collectively. In their introduction to the Chemical Information Ontology (CHEMINF), Hastings et al. explain their case for developing common terminology:[30]

> *Semantic web-enabled software fetches desired data from distributed repositories that support cross-resource query answering over heterogeneous data sources.*

They go on to explain the problems that inhibit federated data-driven research and then describe the rationale and constitution of their CHEMINF ontology.

Conclusions

In broad terms, chemists and other scientists have gained much from the developments of the digital era, not least for interdisciplinary and international collaboration and for the reuse and repurposing of vital data. Regrettably, however, chemists have not taken full advantage of the opportunities provided by electronic methods for improving the preservation, curation, provenance, discovery, and access of research data.

The DataCite organization was formed to encourage data citation and to help researchers to find, access, and reuse data.[31] Interestingly, compliance with DataCite requirements brings into sharp focus the essentials of

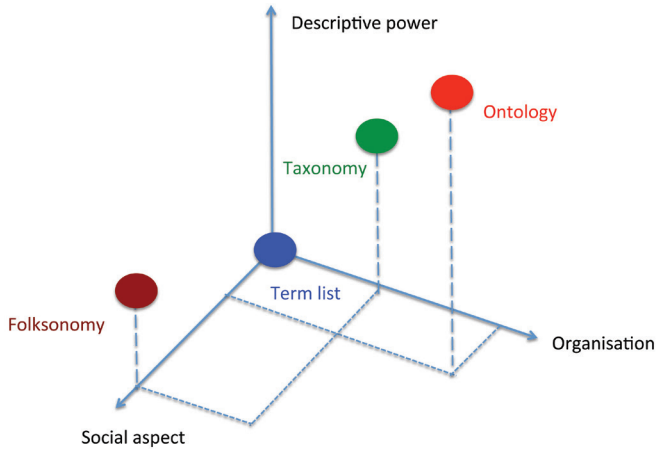

Figure 2: A comparison of the attributes of the principal forms of controlled vocabulary *Reprinted with permission*[3]

curation: institutions that mint their own DOIs undertake to preserve the data and to provide at least some of the essential metadata necessary to discover and access the data, through the medium of a landing page.

Although the advent of DataCite might in itself be a small step, it does motivate researchers to appreciate the value of sharing data and information. To harness and reuse the efforts of others, researchers must acknowledge the necessity and value of curation; they must find innovative ways to overcome the *burden of curation*. This is particularly important for the chemical sciences, if initiatives such as the Dial-a-Molecule Grand Challenge[32] are to be successful. To predict the outcome of novel chemical reactions, extensive information about the results of previous transformations must become available, notably those with unexpected and unsuccessful results. Perhaps it should go without saying that prediction will be impossible unless the data and information relating to previous transformations is satisfactorily curated.

In our own research we have achieved success with automated and mediated capture from instruments and from computations, an area that we intend to develop further. We also consider it essential to stimulate the development of appropriate tools for the capture of provenance and to give particular attention to user interface enhancements that encourage and facilitate the recording of context metadata. Moreover, we consider it essential to promote the agenda for data management education. We take the long-term view that such instruction should begin in secondary and even primary or elementary schools with data management being a basic part of information literacy.

If there is only one message that we leave readers of our paper, it should be to stress the importance of curation at source as one of the fundamental responsibilities of the individual researcher. | FE | doi: 10.3789/isqv25no3.2013.02

COLIN L. BIRD
(colinl.bird@soton.ac.uk) is a visiting researcher in Chemistry and in Electronics and Computer Science.

CERYS WILLOUGHBY
(Cerys.Willoughby@soton.ac.uk) is a PhD student and guest lecturer.

SIMON J. COLES
(s.j.coles@soton.ac.uk) is Senior Lecturer in Chemistry and Director, UK National Crystallography Service.

JEREMY G. FREY
(j.g.frey@soton.ac.uk) is Professor of Physical Chemistry and the UK Digital Economy theme challenge area champion for the IT as a Utility Network+.

Acknowledgments
Our views have been formed both as practicing chemists and information researchers and been brought into focus over the last decade with work 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 (Tender RCUK /D/EPSRC/ Facilities/XRC/10), the HEFCE and JISC Data Management Programme and the University Modernisation Fund (UMF), and most recently the RCUK Digital Economy Theme as part of the IT as a Utility Network+ funding (EPSRC EP/K003569). These views could not have been honed without considerable interaction with our colleagues in Chemistry, Computer Science, and Statistics in Southampton and the e-Research South Consortium (EPSRC EP/F05811X), especially UKOLN, OeRC, STFC, and the DCC, and our professional society and industrial colleagues at the Royal Society of Chemistry, Microsoft Research (MSR), and IBM.

REFERENCES

1. **What is digital curation? [webpage]. Digital Curation Centre (DCC).** www.dcc.ac.uk/digital-curation/what-digital-curation
2. **Bird, Colin L., Cerys Willoughby, and Jeremy G Frey. "Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences."** ***Chemical Society Reviews*****, advance article, July 17, 2013.** http://dx.doi.org/10.1039/C3CS60122F
3. **Bird, Colin L., and Jeremy G Frey. "Chemical information matters: an e-Research perspective on information and data sharing in the chemical sciences."** ***Chemical Society Reviews*****, 2013, 42 (16): 6754–6776.** http://dx.doi.org/10.1039/C3CS60050E
4. **Zeng, Marcia L., and Jian Qin.** ***Metadata*****. New York: Neal-Schuman, 2008.** ISBN: 978-1555706357
5. **Day, Peter.** ***The Philosopher's Tree: A selection of Michael Faraday's writings*****. Bristol, UK: Institute of Physics Publishing, 1999.** ISBN 0750305703
6. **Van Wyhe, John. Darwin Online. [website]** http://darwin-online.org.uk/
7. **Hey, Tony, and Anne Trefethen. "The Data Deluge: An e-Science Perspective." In: Fran Berman, et al., eds.** ***Grid Computing: Making the Global Infrastructure a Reality*****. New York: Wiley, 2003, pp. 809-815.** ISBN: 978-0-470-85319-1
8. **Downing, Jim, Peter Murray-Rust, Alan P. Tonge, Peter Morgan, Henry S. Rzepa, Fiona Cotterill, Nick Day, and Matt J. Harvey. "SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories."** ***Journal of Chemical Information and Modeling*****, 2008, 48 (8): 1571–1581.** http://dx.doi.org/10.1021/ci7004737
9. **Overview of funders' data policies [webpage]. Digital Curation Centre (DCC).** http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
10. **Frey, Jeremy G., and Colin L Bird. "Cheminformatics and the Semantic Web: Adding value with linked data and enhanced provenance."** ***Wiley Interdisciplinary Reviews: Computational Molecular Science*****, Sept./Oct. 2013, 3 (5): 465-481.** http://dx.doi.org/10.1002/wcms.1127
11. **Ince, Darrel. "The Duke University Scandal —What can be done?"** ***Significance*****, September 2011, 8 (3): 113-115.** http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2011.00505.x/abstract
12. **The Reproducibility Initiative [webpage]. Science Exchange.** https://www.scienceexchange.com/reproducibility
13. **Madagascar [wiki]. Reproducible Documents.** http://www.ahay.org/wiki/Reproducible_Documents
14. **Du, Ping, and Joseph A Kofman. "Electronic Laboratory Notebooks in Pharmaceutical R&D: On the Road to Maturity."** ***Journal of Laboratory Automation*****, June 2007, 12 (3): 157-165. :** http://dx.doi.org/10.1016/j.jala.2007.01.001
15. **Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. "Nongeospatial metadata for the ecological sciences."** ***Ecological Applications*****, February 1997, 7 (1): 330-342.** http://dx.doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2
16. **Pancerella, Carmen, John Hewson, Wendy Koegler, et al. "Metadata in the collaboratory for multi-scale chemical science." In:** ***Proceedings of the 2003 international conference on Dublin Core and metadata applications: supporting communities of discourse and practice—metadata research & applications*****. Dublin, OH: Dublin Core Metadata Initiative, 2003, Article #13.** ISBN: 0974530301
17. **Frey, Jeremy. "Curation of Laboratory Experimental Data as Part of the Overall Data Lifecycle."** ***International Journal of Digital Curation*****, 2008, 3 (1): 44-62. :** http://dx.doi.org/10.2218/ijdc.v3i1.41
18. **About CAS [webpage]. Chemical Abstracts Service.** http://www.cas.org/about-cas
19. **Losoff, Barbara. "Electronic Scientific Data & Literature Aggregation: A Review for Librarians."** ***Issues in Science and Technology Librarianship*****, Fall 2009, Issue 59.** http://dx.doi.org/10.5062/F4HH6H0D
20. **Swan, Alma, and Sheridan Brown.** ***The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs*****. JISC, July 2008.** http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf
21. **Digital curation training for all. Digital Curation Centre (DCC).** http://www.dcc.ac.uk/training
22. **Willoughby, Cerys, Colin L, Bird, Emma Tonkin, Simon J, Coles, and Jeremy G. Frey.** ***Metadata in Electronic Lab Notebooks: Experiences with LabTrove*****. In preparation, 2013.**
23. **Borgman, Christine. "Data, disciplines, and scholarly publishing."** ***Learned Publishing*****, January 2008, 21(1), 29-38.** http://dx.doi.org/10.1087/095315108X254476
24. **Beagrie, Neil, Julia Chruszcz, and Brain Lavoie.** ***Keeping Research Data Safe – A Cost Model and Guidance for UK Universities*****. Charles Beagrie Limited, April 2008.** http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf
25. **Boulton, Geoffrey, Philip Campbell, Brian Collins, et al..** ***Science as an Open Enterprise*****. London: The Royal Society, June 2012.** http://royalsociety.org/policy/projects/science-public-enterprise/report/
26. **ChemSpider, The free chemical database. Royal Society of Chemistry (RSC).** http://www.chemspider.com/
27. **Milsted, Andrew J., Jennifer R. Hale, Jeremy G. Frey, and Cameron Neylon. "LabTrove: A Lightweight, Web Based, Laboratory "Blog" as a Route towards a Marked Up Record of Work in a Bioscience Research Laboratory."** ***PLoS ONE*****, July 2013, 8 (7): e67460.** http://dx.doi.org/10.1371/journal.pone.0067460
28. **Frey, Jeremy. "The Value of the Semantic Web in the Laboratory."** ***Drug Discovery Today*****, June 2009, 14 (11-12): 552-561.** http://dx.doi.org/10.1016/j.drudis.2009.03.007
29. **RSC Ontologies [website]. Royal Society of Chemistry.** http://www.rsc.org/ontologies/
30. **Hastings, Janna, Leonid Chepelev, Egon Willighagen, Nico Adams, Christoph Steinbeck, and Michel Dumontier. "The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web."** ***PLoS ONE*****, October 2011, 6(10): e25513.** http://dx.doi.org/10.1371/journal.pone.0025513
31. **DataCite: Helping you to find, access, and reuse data.** http://www.datacite.org/
32. **Dial-a-Molecule EPSRC Grand Challenge Network.** http://www.dial-a-molecule.org/wp/