Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Chapter 9

Preparing To Support Research Data Sharing

Ye Li* **and Lori Tschirhart**

**Shapiro Science Library, University of Michigan,
Ann Arbor, Michigan 48109
*E-mail: liye@umich.edu**

When national funding agencies introduced data management requirements for grant proposals, details were scant and researchers turned to research-supporting staff for assistance with compliance. Support staff has experienced a sudden demand for e-Science and data sharing knowledge and expertise, often before institutional infrastructures and strategic plans have been developed. As a part of the research-supporting system, we share our learning paths, resources, and strategies here. We also describe and analyze emerging needs in the Chemistry domain to demonstrate discipline-specific data sharing issues and approaches used to customize services for local research communities.

Introduction

When we started our current positions as subject specialists at the University of Michigan's Shapiro Science Library in 2009, we had not anticipated how quickly e-Science, data sharing and data management would become central aspects of our job. What started with the 2003 Atkins Report by the National Science Foundation (NAF) Blue-Ribbon Advisory Panel on Cyberinfrastructure, which declared the need and potential for an e-Science revolution for science and engineering research in the U.S. (*1*), ended with funding agencies including and mandating data management components to their guidelines. In between, there were several important articles published on the subject (e.g., Anna K. Gold's two-part article (*2*, *3*)), as well as a 2009 guide discussing how subject specialists could collaborate with researchers on e-Science projects at Purdue University (*4*). Now there are many other reports from NSF, other national agencies, and related organizations that provide an overview of cyberinfrastructure and e-Science in the

© 2012 American Chemical Society
In Special Issues in Data Management; Xiao, Norah, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2012.

U.S. and around the world. Some selected documents are linked on our research guides at http://guides.lib.umich.edu/ci. The reports clarify the big picture and sometimes present domain-specific research needs and challenges.

However, when funding agencies were only just beginning to include and mandate data management components in their guidelines, paths to meeting the guidelines were not always obvious. The burden was even more challenging for fields like Chemistry since it is considered a "small science," a "long-tail science," and a somewhat proprietary discipline without a data sharing tradition (*5*). A recent publication, *The fourth Paradigm: data-intensive Scientific Discovery* (*6*), highlights the research potential for those domains not yet obvious part of "big data." For example, chemistry as a basic science domain plays an important role in all the "big data" research areas. Though many chemists running individual laboratories have not yet seen the direct value of sharing their data, we are beginning to see movement from academia, cooperatives, and publishers.

As data management plans became mandated through funding agencies, researchers turned to research-supporting staff for quick solutions to address the requirements. The research-supporting staff to which the researchers turned consisted of grant officers, institutional repository (IR) service providers, information scientists, graduate students in research groups, and, of course, librarians. Although our librarian job titles and position descriptions did not suggest data management responsibilities, we participated in finding solutions for many reasons: part of our mission is to support institutional researchers; our existing relationships with institutional researchers provide us with an awareness of researchers' data management needs and wants within areas of disciplinary expertise; and our membership in a profession with a tradition of collecting, managing, storing, and making accessible other types of research output affords us valuable insights.

Therefore, we found motivation to engage with campus researchers and began to help with the data storage and sharing needs of researchers at our institution. Importantly, we had administrative encouragement to pursue conversations and discover channels to gain knowledge. With this license, we developed strategies taking advantage of librarians' expertise, campus expertise, and expertise around the world. Now, as our institution is still building our own infrastructure for e-Science, we offer an account of our learning journey including obstacles encountered, resources, and expertise we drew upon, and our approaches to understanding the research community we serve. Our perspective may also help staff from the other research-supporting groups who are just beginning this process to prepare themselves to help meet data sharing needs.

Identify Learning Tools

Once we understand the fundamentals of e-Science and cyberinfrastructure principles, we are ready to learn how to match our own expertise with what researchers need in the disciplinary domains that we support. Since data science is a rather new research domain, foundational literature is not as abundant as in

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

other domains. Relevant research articles are regularly published and should be read, but many other learning channels and learning strategies are available now.

Organizations and Their Publications

Since the emergence of e-Science and cyberinfrastructure, many organizations have dedicated study to data-intensive research issues. Some provide repositories, software, frameworks or other data related tools; some are specially funded organizations dedicated to the evolution of new tools, frameworks and services. Regardless of organizational scope, publications from these organizations are often explicitly shared and can be used as direct learning tools. Earlier we mentioned reports from government agencies and other related organizations for overviews and strategic planning. Here, we list some examples of dedicated research organizations.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

- Digital Curation Center (DCC, http://www.dcc.ac.uk/) As the leading center of the United Kingdom's effort on research data management strategy and practice development, DCC hosts a rich open access collection of project documentation, standards, case studies, tutorials and other training documents (*7*). The topics cover issues most crucial to data sharing and data management. Although the amount of information here may be overwhelming for beginners, we still recommend it as the top resource to consult early in the process.
- Inter-University Consortium for Political and Social Research (ICPSR, http://www.icpsr.umich.edu) With 50 years of experience working with social science data, ICPSR is a leader in data management within and beyond the social science domain. ICPSR is concerned with whole lifecycle data curation, analysis, and access. The consortium sponsors data science research and instruction related to data and maintains a robust data repository. The Digital Curation section (*8*) and the Guidelines for Effective Data Management Plans section (*9*) on the ICPSR website are particularly helpful with practical considerations in data management. While the materials are prepared to support Social Science research, research-supporting staff in other research domains will also gain a systematic view of issues to be considered for data management and sharing.
- DataOne (http://www.dataone.org/) and Data Conservancy (http://dataconservancy.org/) The two projects were funded by NSF in 2009 for different focuses – DataONE is devoted to building a framework, cyberinfrastructure, and data repository for environmental science and related fields; the Data Conservancy develops software for data repositories, explores data sharing practices, and fosters development of community, tools, and services for data re-use across social science and science disciplines. Publications on DataOne (*10*) and Data Conservancy (*11*) provide good reference articles for future projects; so do the publications listed on the Dataverse Network Project (*12*). With these publications, individuals can find papers documenting details involved

in the process of making data reusable and translate those details into the domains they support. These resources and tools provided on these sites may also be shared with institutions and researchers.

- Association of Research Libraries (ARL, http://www.arl.org/) For the librarian community, the Association of Research Libraries is concerned with policy and scholarly communication issues that impact libraries. Because e-Science will influence the way scholars communicate and because policy decisions direct the development of cyberinfrastructure, ARL is exploring what librarians can do for e-Science. For example, the ARL study of member institution activity with e-Science and data services provides a sketch of how libraries started services in this area (*13*). What individual subject specialists and other librarians can do in practice is also explored by the members of the ARL e-Science task force (*14*).

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Conferences and Workshops

Data science conferences and workshops gather people wishing to communicate recent work, learn from each other, get inspired, and generate and apply new ideas related to data research. Direct communication with peers supporting research data sharing is an efficient way to learn concepts and the research interests of fellow attendees.

The International Digital Curation Conference hosted by DCC (http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc) is one of the largest international events for data science and practice. Whatever your focus or niche, you may find peers working on similar topics at this conference. Presentations and videos from the events are also available on DCC website.

If data sharing is important to a domain, then sessions dedicated to the topic wil likely be found at major conferences for that domain. Librarians may also find valuable programming within the information divisions of domain-specific conferences and within the domain divisions of library association conferences. For example, relevant panels and oral presentation sessions have been organized by Chemical Information Division (CINF) of the American Chemical Society (ACS) at the ACS National Meetings, and by the Division of Chemistry (DCHE) of the Special Library Associations (SLA) at the SLA Annual Conference and Info-Expo for the past three years.

Various institutions and organizations have realized the importance of "train-the-trainers" events and offer workshops to share their expertise in data sharing.

One example is the Data Curation Profile (DCP) Toolkit workshop provided by D. Scott Brandt and Jacob R. Carlson from the Purdue University Libraries (*15*). The workshop series is funded by the Institute of Museum and Library Services (IMLS) to train librarians to interview discipline-appropriate researchers and populate a DCP repository. The accumulated DCPs can be used to reveal the data curation needs of different research communities. Attendees learn how to complete a DCP while thinking through the whole data lifecycle and associated curation issues. Although not all attendees will have the opportunity to execute extensive interviews with researchers, the tool kit provides a framework to

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

organize data-sharing conversations with researchers and successful interviews may help to demonstrate the research-supporting commitment of the librarians conducting the interviews.

ICPSR also provides train-the-trainer style workshops. The ICPSR Summer Program has provided data processing and management training for social scientists around the world since 1963 (*16*). Recently, ICPSR expanded training opportunities with an "Applied Data Science: Managing Research Data for Re-Use" workshop, designed to provide a platform for sharing the expertise from ICPSR, University of Michigan, and all around the world. The workshop combines a big picture overview, detailed case studies, a resource summary, and research updates for the data science field. The balance between discussion of practical issues and up-to-date exploration is the strength of this workshop.

Organizations such as DCC and ICPSR are starting to provide online training programs. While some participants may miss the face-to-face experience that comes with physically attended programs, online training programs offer convenience. Hybrid programs are also emerging such as Data Intelligence 4 Librarians. Developed by 3TU.Datacentrum at Netherlands (http://dataintelligence.3tu.nl/en/home/), the education course provides a mix of online and group meeting learning opportunities. We expect that similar programs will emerge in the U.S. in the near future.

Personal and Organizational Communications

Formal and informal communication with peers can be useful for exchanging problems and stories. When shared, the knowledge and experience of peers can orient research-supporting staff to relevant data sharing issues. Practices that addressed earlier problems may be applied across disciplines and institutions to resolve current problems. Peers are often the best sources of information for surveys and reports previously conducted within our institution. Often, consultation with colleagues is the best way to discover a local institution's history of data management exploration. For instance, a campus wide survey regarding researchers' data management practices conducted by the School of Information at the University of Michigan in 2010 (*17*) was serendipitously discovered by librarians in July 2012 while attending a presentation delivered by the survey author. Compiling such hard-to-find works promotes additional discovery.

RSS feeds and listservs often point to helpful resources and tools. Listservs also provide convenient forums for questions and discussions with peers. Listservs of particular benefit for chemical information specialists include the CHMINF-L (*18*) and SLA-DCHE (*19*). The listserv of Office from the Research Cyberinfrastructure at the University of Michigan (*20*) is crucial for keeping up with activities within our local institutional organization.

Searching online content sharing and social network platforms can also lead to valuable learning channels. Videos, tutorials, and project presentations about e-Science and data sharing are available in abundance via Youtube and Slide Share websites. Following tweets of specific events on data topics and by distinguished researchers in data science on Twitter helps users to stay current. Maintaining a refined list of RSS feeds from interested organizations is another way to keep up to date. For example, RSS feed from the NSF Office of Cyberinfrastrcture Discoveries (http://www.nsf.gov/rss/rss_www_discoveries_oci.xml) provides current information about cyberinfrastructure-enabled discovery.

The tools and channels we mentioned above are those we have used. Beginners are encouraged to explore whichever resources best fit their needs. Resources created for distinct research communities hold significant value for unintended research domains and should not be ignored. In this early era of e-Science, inspiration and critical information may be found in any relevant research and discussions, regardless of the research community for which it was originally intended.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Enrich Your Toolbox

The learning tools and resources described above may be used directly by researchers to manage and inform their data sharing practices and by research-supporting staff to communicate with researchers. Since any of those resources could be more helpful in some contexts over others, it is useful to collect and organize these resources and tools into a toolbox. Table 1 provides a summary of necessary tools and resources. Depending on the needs of the research community, one can consult different resources and tools to address an immediate need while furthering the personal learning process.

Table 1. Tools and resources to be collected to support data sharing

Tools and Resources	*Examples*
Institutional policy about research data	Research policy pages of various institution(s)
Data management plan (DMP) templates	DMP Online (*21*), DMP Tool (*22*), DMP templates by various universities
List of disciplinary repositories	Databib (*23*), OAD: Data repositories (*24*), Data Cite: Repositories (*25*)
Profiling/communication tools	DCP Toolkit (*26*)
Institutional repositories	DSpace@MIT (*27*), DataStaR (*28*), PURR (*29*)
Metadata standards available	DDI Metadata resources (*30*), Science Data Literacy Project: Metadata Standards (*31*), Metadata standards and related resources on D2I Wiki (*32*), CML (*33*)
Data citation	DataCite (*34*), ICPSR: Data Citation (*35*)
Teaching materials for data literacy	Science Data Literacy Project (*36*), e-Science Portal for New England Librarians: Science Data Literacy (*37*)

Besides the tools and resources listed in Table 1, any well-written documents that describe best research/data practices for a relevant research community should be collected and offered up as recommended reading for researchers. Research-supporting staff has opportunities to help research communities improve and perfect their research practices.

Sometimes, the set of tools and resources will not be ideal for certain research community needs. Research-supporting staff must work with researchers and existing tools to develop new, custom tools to fit researchers' needs and disciplinary data re-use needs.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Focus on an Individual Research Community

To identify and prioritize the research-supporting services most essential to a research community, to communicate local research community priorities to institutional stakeholders and leaders, and to make meaningful contributions to infrastructure development, we need to truly analyze the domains and research communities that we serve. Here, we use the domain of Chemistry and the research community connected with the Department of Chemistry at the University of Michigan as examples to demonstrate our approaches.

As mentioned in the introduction to this chapter, some characteristics of the chemistry domain pose unique data sharing problems. The "small science" nature of the domain emphasizes research conducted individually or in small groups. This characteristic may limit researchers' perceptions of the utility and potential of large-scale data sharing within the domain. Although many sub-disciplines of Chemistry focus on individual lab works and traditional publications, an exception exists in the area of crystallography where crystal structures are often deposited to the Cambridge Structural Database (CSD) (*38*) and are curated by staff at the Cambridge Crystallographic Data Center (CCDC). Although direct mining of the crystal structure data can still be challenging due to missing metadata, crystallography is farther along in the data-sharing universe than most other areas of Chemistry.

The relatively proprietary nature of chemistry research poses another barrier to data sharing, since those with ownership stakes in chemistry research stand to lose even as society gains from large-scale data sharing. However, unsustainable increases in research and development cost have prompted increased collaboration between pharmaceutical companies and academic researchers in all stages of their drug development programs (*39*). These collaborations may loosen access to the long-locked gates to some internal databases of those pharmaceutical companies, expedite new drug developments, and reduce R&D cost for pharmaceutical companies eventually. The benefits of large-scale data sharing may then be recognized by stakeholders of this industry.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Sources of Data, Data Types, and Research Profiles in Chemistry

Chemistry data can be found in a variety of sources. Major providers of small molecule data sources are summarized in Table 2. Most Chemistry data generated from research in academia are presented in publications such as journal articles and their supplemental materials, which are then indexed and made searchable in databases and reference books. These databases are relied upon frequently by chemists. Unfortunately, much of the data published in the literature are presented in formats that don't allow for re-use and are provided without associated metadata. Data re-use here is also limited by the high cost of access to the proprietary databases which make data discoverable.

Table 2. Major types of sources for data of small molecules in chemistry

Source	*Examples*	*Metadata*	*Re-use*
Publications	Journal articles and supplemental materials	Buried in texts and captions	No
Proprietary databases indexing data	CAS databases, Reaxys, ASM Phase Diagrams	Limited	Possible but locked up
Reference books indexing data	CRC Handbook of Chemistry and Physics, Springer Materials	Limited	Mostly not
Drug screening databases	ZINC, Internal databases in pharmaceutical companies	Some	Possible except proprietary ones
Open access "hybrid" databases *	ChemSpider, PubChem	Some	Possible
Disciplinary repository	Cambridge Structural Database	Yes	Possible
Institutional repository	DatastaR, PURR	Some, not specific for Chemistry data	Possible

* Content comes from both data mining and users depositing

Chemical Abstract Services (CAS) databases (*40*) and Beilstein database (now a part of Reaxys (*41*)) have existed (first as print, then as electronic databases) for over a hundred years and cover literature back to 18th century. These databases provide a valuable service by extracting data from static publications and making them discoverable. The original purpose of these databases was to make the data and associated publications discoverable, but not reusable, either by human or machine. Technology may offer new potential for content re-use within these rich collections if the issue of costly access could be resolved. Reference books index data similarly to these databases. Databases like ZINC (*42*), which is designed for drug screening, have the most "big data" re-use potential. Significant access barriers exist for most other drug-screening databases containing experimental data due to proprietary interests of large pharmaceutical companies while free databases like ZINC contain mostly data from theoretical calculation.

The potential for data re-use is strongest within the last three sources listed in Table 2. Open access hybrid databases like ChemSpider (*43*) and PubChem (*44*) are designed with different research communities in mind but both emphasize work with small molecules. These two resources share a common strength in policies allowing the public to deposit data. A large amount of work must be dedicated to the curation of the publicly-deposited data before it becomes truly reusable, so resources must be allocated to that curation. Disciplinary and institutional repositories are still in their early stages of development and use. Repository developers are exploring reasonable preservation and access models, especially with regards to metadata standards, to allow effective re-use of chemistry data in these repositories.

A brief discussion of the data of small molecules does not represent the interdisciplinary nature of current Chemistry research and it does not address the importance of the data of polymers and biological molecules in the domain. We expect that future disciplinary repositories will be based upon research themes, such as Energy Science, more often than upon traditional disciplines like Chemistry, because data types and purposes of re-using data are more homogeneous within the same research theme than those across the traditional disciplines. In addition, researchers working on similar research questions tend to participate in active research communities and will be inclined to deposit data where it will be most useful to those working on answering similar research questions.

To illustrate the heterogeneity we have in traditional disciplines like Chemistry, we present a non-exhaustive representation of sample data types associated with Chemistry in Table 3.

Based on the sample data types presented in Table 3, a repository inclusive of all data types that could accommodate preservation and re-use requirements would pose extraordinary problems related to metadata standardization, accessibility, and interoperability. The challenge to create a uniform metadata standard for all the data types here, especially for the metadata describing provenance and experimental conditions, demonstrate the problems associated with this type of repository.

Universities and other institutions are trending toward grouping their researchers by research themes within or beyond traditional departments. The Chemistry Department at the University of Michigan is presented as one example. Figure 1 depicts how principal investigators (PIs) are distributed in both the traditional disciplines and research themes. The data used to plot Figure 1 are summarized from https://www.chem.lsa.umich.edu/chem/faculty/research.php in July 2012. From the bottom graph (based on research themes) in Figure 1, we can see that the majority of PIs have research focuses related to biochemistry, energy, and imaging. PIs often have interest in multiple research themes, which implies that data collected in his/her lab may be of interest to multiple research communities. When we think about data curation for the data produced by these

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

research themed groups, we must consider the data needs of intersecting research communities.

Table 3. Sample small molecule data types in chemistry

Sub-domain	*Example Data Types*
Synthetic Chemistry	
Preparation procedure	Text with chemical names and special symbols, Scheme
Substance	Identifier, Structure
Characterization/purification	
Spectroscopy	1D and 2D NMR/IR/Mass Spectra, UV-Vis spectra, Atomic Absorption Spectra, Fluorescence Spectra, Raman Spectra
Numerical Data	Boling point, melting point, solubility, etc.
Chromatography	HPLC, GC, CE
Crystallography	Crystallographic structure, Crystal preparation
Computational Chemistry	Gaussian log files
Microscopy	Photomicrograph SEM Image/Video TEM Image/Video AFM Image/Video Confocal microscopy Image/Video
Electrochemistry	Standard electrode potential, Resistance, Voltammetry, Coulometry
Physical chemistry	
Thermodynamics	Entropy, enthalpy, etc.
Kinetics	Reaction rates etc.
Surface chemistry	Adsorption coefficient, etc.

To understand which data types are generated most frequently within our Chemistry Department, we run an ongoing study to profile data types found within publications authored by PIs in the Department. Publications meeting certain criteria are retrieved from the Web of Science index, grouped by PI, and made into a reference set. A FileMaker database was established to hold information extracted from the reference set, including bibliographical information and occurrences of data types that appeared in figures, tables, texts, captions, and especially supplement materials. A controlled vocabulary is in development to describe the data types in consultation with researchers in the Chemistry Department.

The results of this data profiling study will be reported separately. Here are some preliminary statistics about the reference set we collected: 635 journal

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

articles and edited book sections published between 2010 and May 2012 are authored by PIs in the Department. 51 of these articles represent collaborations between two PIs and four are written among three PIs. Ten out of the twelve journals in which our PIs published most are ACS Publications, which means that the availability of data from these publications is highly dependent on the publishing practice of ACS Publications. If publishers like ACS Publications encourage or require data publishing associated with articles as they have done for crystallography data, researchers will have additional incentive to integrate data publishing as a part of scholarly communication thus fostering data sharing.

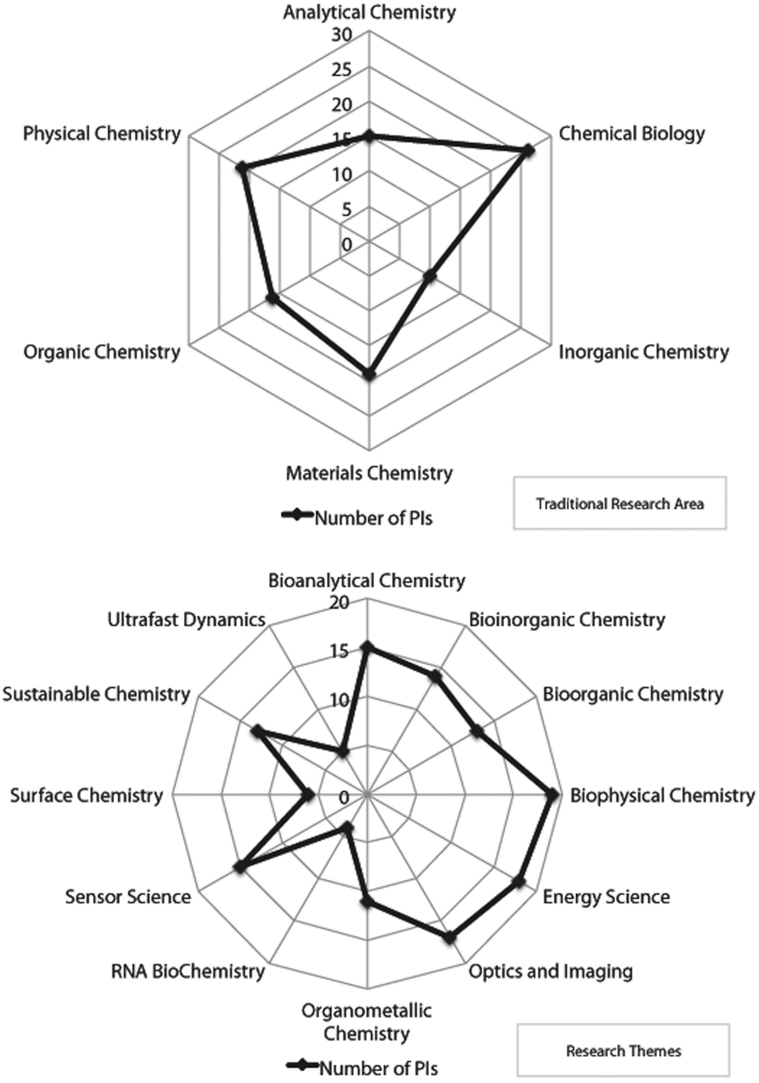

Figure 1. Research Profiles of Chemistry Department at the University of Michigan in Traditional Research Areas and in Research Themes.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

In this instance, data types were profiled through publication analysis as an alternative to conducting Data Curation Profile (DCP) interviews with PIs in the Chemistry Department. Our approach reveals the overall departmental profile more directly while DCP provides the complete data story for an individual research projects. Given sufficient time and resources, a combination of the two approaches will provide perspectives at both macro and micro levels.

Metadata

Metadata and metadata standards are crucial for data preservation and sharing. To ensure proper long-term preservation, accessibility and reusability, we need a minimum of descriptive, administrative, and structural metadata (*45*).

Below, we examine a few current disciplinary repositories used for Chemistry data to see how data formats and metadata needs are handled. The results are summarized in Table 4.

As shown in Table 4, three characteristics persist among the examined repositories: (1) data types are limited to crystallography, spectra, structure and some reaction data; (2) data formats are not always suitable for long-term preservation; (3) the amount of metadata required for deposit is minimal and limited to bibliographic and technical metadata. Despite the room for improvement, some encouraging trends are emerging. Two of the repositories request description of reaction conditions and experimental details. These descriptions may be annotated with markup language like XML and become machine-readable descriptive metadata. ChemSpider also asks for explanation of characterization data, which can be annotated into structural metadata to show the relationship among the characterization data and the identifiers. Finally, elective embargo periods are becoming important components of the administrative data for a couple of the repositories. In all, it seems that repositories are lowering barriers to deposit by requesting minimal metadata from depositors. It may be a good strategy to jumpstart repository population, but more systematic collection of metadata and better metadata standards will benefit more data sharing long-term. This strategy is consistent with what Jane Greenberg *et al* described as best practices for a scientific data repository in their 2009 publication (*46*) based on practices of Dryad repository, which was designed for evolutionary biology, ecology, and related disciplines.

In Chemistry, Chemical Markup Language (CML) was an early success for describing data for the semantic web (*33*). Software such as the Microsoft Chem4Word has integrated CML in the package (*47*). Currently, CML supports molecules, compounds, reactions, spectra, crystals and computation chemistry.

Although CML is not designed to be a metadata standard for data repositories, it is an excellent candidate to become a standard for data repositories in Chemistry. In fact, the first two repositories listed in Table 4, eBank-UK eCrystal and SPECTRa have already used subsets of CML to encode metadata and allowed direct export of metadata as CML files.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Table 4. Data format and requested metadata elements of selected repositories/projects in chemistry

Project/ Repository	*Data type explored*	*Standardized data format*	*Metadata requested when depositing*
eBank-UK eCrystal (*48*)	Crystallography	CIF, HKL files	• Bibliography • Data collection parameters • Stages of the structure determination • Experimental conditions
SPECTRa (*49*)	Crystallography	CIF files	• Extracted from CIF, JCAMP, and Gaussian file • Bibliography • Embargo period
	NMR	JCAMP-DX and MDL mol files	
	Computational Chemistry	Gaussian Archive files	
Cambridge Structural Database (*38*)	Crystallography	CIF, FCF or HKL files	• Extracted from CIF • Bibliography • Associated publication • Keywords about study
ChemSpider (*43*)	Chemical structure	MOL, SDF, CDX, SKC files	• General description • Identifiers • Links to websites or publications
	Spectra (^{1}H NMR, ^{13}C NMR, IR and Mass)	JCAMP-DX or –JDX files JPG or PNG for 2D NMR	• Extracted from JCAMP file • Link to associated webpage • Experimental details in Comments field
	Synthetic reaction and associated characterization data	TXT , ChemDraw ChemSketch or RXN file as well as GIF or PNG file for Scheme	• Bibliography • Embargo period • Chemicals involved • Link to publications • Experimental details in Comments and Multimedia fields • Explanation about characterization data • Reaction keywords

Many other markup languages (*50*), such as ThermoML, AnIML, UnitsML have the potential to be used together with CML to create proper metadata standards in Chemistry. We should note that these markup languages are only useful for descriptive and some structural metadata of data in Chemistry. Repositories still need to amend administrative and structural metadata in practice, especially those related to regulation compliance and privacy protection. The creation of a master suite of metadata standards for major data types in Chemistry would benefit localized metadata standards. Various repositories built around different research themes could adopt subsets of the master standards to use in combination with locally developed metadata requirements.

What Do Researchers Really Need?

Regardless of research community, the following questions help research-supporting staff understand the data sharing needs of their constituents. What is the research profile of the department? What types of data are important? What metadata are necessary for various research themes and data types in the department? How can we identify potential data consumers, the designated community, for data from this community? Are there existing repositories to be recommended to our researchers? Will depositing into an institutional repository support this community? How do research practices in the department influence data sharing? Can research-supporting staff help to improve the research workflow so that data sharing becomes effortless and effective? Answers to these questions may take years to find. One urgent question, however, begs to be answered right now: what do researchers really need? This question can be addressed in two means: one practical and one ideal.

The practical approach examines the whole research lifecycle from idea formulation to proposal writing, project planning, data collection, data processing, publishing and sharing, and back to idea formulation to see which parts of the process have been supported by institutional facility and personnel. Meanwhile, we need to consider how the data lifecycle, the data curation cycle, and the scholarly communication cycle can be integrated with the research lifecycle. If any areas not currently being served are identified in the cycles or the integration process of the cycles, these gaps are where future services should be focusing on. This approach has been described by Jacob Carlson from Purdue University at an ICPSR workshop in July 2012. One advantage of this model is that it can be used for institutional strategic planning and also can be used by research-supporting staff to prioritize tasks to support research communities. The model also allows for improvements to the quality of research-supporting services as a whole instead of narrowly focusing on data related tasks.

Based on our own research experience and communication with researchers, the ideal world would involve highly automated workflows enabled by sophisticated lab management systems. Using such systems, any meaningful activities in the lab, from idea being generated, to experiment process, data processing, and paper writing, would be facilitated, recorded and curated with semantic annotations. Then, the content can be selectively and directly shared with designated communities. Backup and preservation of all content would happen behind the scenes without extra effort made by researchers. Since the system and the workflows would be standardized and interoperable, anyone with a need to re-use the data could precisely extract the shared data. Everything shared would be shared with context and recorded provenance creating an ideal environment for data re-use. Labs around the world would essentially be one lab with different rules for different components. The ideal world is far away but possibilities are already emerging. In the domain of Chemistry, a series of projects, including CombeChem, Smart Tea, R4L, e-Bank and e-Crystals, are

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

led by a group of UK scientists (*51*) to create lab management systems similar to what we described. Components of these systems are in development under the umbrella of the Smart Research Framework (SRF) collaborative systems (http://www.mylabnotebook.ac.uk/). Technology revolutions may enable us to realize this ideal research world sooner than we can imagine. Assisting researchers to cultivate good lab practices with the data-centered paradigm in mind will prepare them for the exciting new era in Science.

Learn, Teach, and Collaborate Simultaneously

As a part of teaching and research supporting system of universities, librarians are simultaneously learning and teaching new knowledge and skills as well as collaborating with faculty, students, and staff across campus to accomplish various projects. The emergence of e-Science and data sharing is an opportunity for us to provide new services through the same means. We can apply our expertise in organizing, archiving, and preserving information as well as our traditional roles as connectors of different disciplines on campus. We are nurturing our new expertise in supporting the research cycle, data lifecycle, scholarly communication cycle, and curation of all scholarly processes and outputs. We hope our shared experiences here orient and inspire beginners to get started with this exciting exploration.

Acknowledgments

We would like to thank the University Libraries at the University of Michigan for all of the learning opportunities provided to us. Thanks also to our colleagues, faculty, students, and staff at the University of Michigan for their support for our work.

References

1. Atkins, D. E.; Droegemeier, K. K.; Feldman, S. I.; Garcia-Molina, H.; Klein, M. L.; Messerschmitt, D. G.; Messina, P.; Ostriker, J. P.; Wright, M. H. *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*; Office of Cyberinfrastructure National Science Foundation, January 2003.
2. Glod, A. Cyberinfrastructure, Data, and Libraries. Part 1: A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine*, 2007. http://www.dlib.org/dlib/september07/gold/09gold-pt1.html (accessed July 2009).
3. Gold, A. Cyberinfrastructure, Data, and Libraries. Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries. *D-Lib Magazine*, 2007. http://www.dlib.org/dlib/september07/gold/09gold-pt2.html (accessed July 2012).

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

4. Garritano, J. R.; Carlson, J. R.. A Subject Librarian's Guide to Collaborating on e-Science Projects *Issues in Science and Technology Librarianship*, Spring 2009. http://www.istl.org/09-spring/refereed2.html.
5. Velden, T.; Lagoze, C. Communicating chemistry. *Nat. Chem.* **2009**, *1* (9), 673–678.
6. Hey, A. J. G.; Tansley, S.; Tolle, K. M. *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Microsoft Research: Redmond, WA, 2009.
7. Resources for Digital Curators. http://www.dcc.ac.uk/resources (accessed July 2012).
8. ICPSR: Digital Curation. http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/dmp/index.html (accessed July 2012).
9. ICPSR: Guidelines for Effective Data Management Plans. http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/dmp/index.html (accessed July 2012).
10. DataONE Publications. http://www.dataone.org/publications (accessed July 2012).
11. Data Conservacy: Browse Our Collection of White Papers and Publications. http://dataconservancy.org/library/ (accessed July 2012).
12. Dataverse Network Project: Publications. http://thedata.org/publications (accessed July 2012).
13. Soehner, C.; Steeves, C.; Ward, J. *E-Science and Data Support Services: A Study of ARL Member Institutions*; Association of Research Libraries: Washington, DC, 2010.
14. Gabridge, T. The last mile: Liaison roles in curating science and engineering research data. *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC* **2009** (265), 15–21.
15. DCP Toolkit Workshops. http://datacurationprofiles.org/workshops_content (accessed July 2012).
16. Summer Program in Quantitative Methods of Social Research. http://www.icpsr.umich.edu/icpsrweb/sumprog/index.jsp (accessed July 2012).
17. Fear, K. "You Made It, You Take Care of It", Data Management as Personal Information Management. *Int. J. Digital Curation* **2011**, *6* (2), 53–77.
18. Chemical Information Sources Discussion List. http://www.lsoft.com/scripts/wl.exe?SL1=CHMINF-L&H=LISTSERV.INDIANA.EDU (accessed July 2012).
19. SLA Chemistry Division: DCHE Listserv. http://chemistry.sla.org/listserv/ (accessed July 2012).
20. University of Michigan Office of Research Cyberinfrastructure: News + Events. http://orci.research.umich.edu/news-events/ (accessed July 2012).
21. DMP Online. https://dmponline.dcc.ac.uk/ (accessed July 2012).
22. DMP Tool. https://dmp.cdlib.org/ (accessed July 2012).
23. Databib. http://databib.org/index.php (accessed July 2012).
24. Open Access Directory (OAD): Data Repositories. http://oad.simmons.edu/oadwiki/Data_repositories (accessed July 2012).
25. DataCite: Repositories. http://www.datacite.org/repolist (accessed July 2012).

26. Data Curation Profile Toolkit. http://datacurationprofiles.org/ (accessed July 2012).
27. DSpace@MIT. http://dspace.mit.edu/ (accessed July 2012).
28. DataStaR. http://datastar.mannlib.cornell.edu/ (accessed July 2012).
29. Purdue University Research Repository (PURR). https://research.hub.purdue.edu/ (accessed July 2012).
30. DDI Medtadata Resources. http://www.ddialliance.org/metadata-resources (accessed July 2012).
31. The Science Data Literacy Project: Metadata Standards. http://sdl.syr.edu/?page_id=32 (accessed July 2012).
32. Data to Insight Center: Matadata Standards and Related Materials. http://d2i.indiana.edu/wiki/Metadata_Standards_and_Related_Materials (accessed July 2012).
33. Chemical Markup Language. http://www.xml-cml.org/ (accessed July 2012).
34. DataCite. http://www.datacite.org/ (accessed July 2012).
35. ICPSR: Data Citation http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/citations.jsp (accessed July 2012).
36. The Science Data Literacy Porject : Educator Resources. http://sdl.syr.edu/?page_id=15 (accessed July 2012).
37. e-Science Portal for New England Librarians: Science Data Literacy. http://esciencelibrary.umassmed.edu/sci_data_literacy (accessed July 2012).
38. Cambridge Structural Database (CSD). http://www.ccdc.cam.ac.uk/products/csd/ (accessed July 2012).
39. Coles, L. D.; Cloyd, J. C. The role of academic institutions in the development of drugs for rare and neglected diseases. *Clin. Pharmacol. Ther.* **2012**, *92* (2), 193–202.
40. Chemical Abstract Services (CAS) Databases. http://cas.org/expertise/cascontent/index.html (accessed July 2012).
41. Reaxys. https://http://www.reaxys.com/info/ (accessed July 2012).
42. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52* (7), 1757–1768.
43. ChemSpider. http://www.chemspider.com/ (accessed July 2012).
44. PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed July 2012).
45. Green, A.; Macdonald, S.; Rice, R. *Policy-Making for Research Data in Repositories: A Guide*; Data Inormation Specialists Committee-UK: U.K., 2009.
46. Greenberg, J.; White, H. C.; Carrier, S.; Scherle, R. A metadata best practice for a scientific data repository. *J. Libr. Metadata* **2009**, *9* (3−4), 194–212.
47. CML-Aware Software. http://www.xml-cml.org/tools/software.html (accessed July 2012).
48. Coles, S. J.; Frey, J. G.; Hursthouse, M. B.; Light, M. E.; Milsted, A. J.; Carr, L. A.; DeRoure, D.; Gutteridge, C. J.; Mills, H. R.; Meacham, K. E.; Surridge, M.; Lyon, E.; Heery, R.; Duke, M.; Day, M. An e-science environment for service crystallography: From submission to dissemination. *J. Chem. Inf. Model.* **2006**, *46* (3), 1006–1016.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009

49. Downing, J.; Murray-Rust, P.; Tonge, A. P.; Morgan, P.; Rzepa, H. S.; Cotterill, F.; Day, N.; Harvey, M. J. SPECTRa: The deposition and validation of primary chemistry research data in digital repositories. *J. Chem. Inf. Model.* **2008**, *48* (8), 1571–1581.

50. Nic, M. Chemical XML Formatting. In *Chemical Information Mining: Facilitating Literature-Based Discovery*; Banville, D. L., Ed.; CRC Press: Boca Raton, FL, 2009; pp 99−119.

51. Frey, J. Curation of laboratory experimental data as part of the overall data lifecycle. *Int. J. Digital Curation* **2008**, *3* (1), 44–62.

Downloaded by MICHIGAN STATE UNIV on January 21, 2016 | http://pubs.acs.org
Publication Date (Web): November 15, 2012 | doi: 10.1021/bk-2012-1110.ch009