J Comput Aided Mol Des (2014) 28:975–988
DOI 10.1007/s10822-014-9777-4

Academic librarians at play in the field of cheminformatics: building the case for chemistry research data management

Leah McEwen · Ye Li

Received: 17 April 2014 / Accepted: 7 July 2014 / Published online: 20 July 2014
© Springer International Publishing Switzerland 2014

Abstract There are compelling needs from a variety of camps for more chemistry data to be available. While there are funder and government mandates for depositing research data in the United States and Europe, this does not mean it will be done well or expediently. Chemists themselves do not appear overly engaged at this stage and chemistry librarians who work directly with chemists and their local information environments are interested in helping with this challenge. Our unique understanding of organizing data and information enables us to contribute to building necessary infrastructure and establishing standards and best practices across the full research data cycle. As not many support structures focused on chemistry currently exist, we are initiating explorations through a few case studies and focused pilot projects presented here, with an aim of identifying opportunities for increased collaboration among chemists, chemistry librarians, cheminformaticians and other chemistry professionals.

Keywords Chemistry librarians · Research data management · Chemistry metadata · Chemical health and safety

L. McEwen (✉)
Physical Sciences Library, Cornell University, 283 Clark Hall, Ithaca, NY 14853, USA
e-mail: lrm1@cornell.edu

Y. Li
Shapiro Science Library, University of Michigan, 919 South University Avenue, Ann Arbor, MI 48109, USA

Introduction

During the past decade, everywhere we hear a great call for scientific data, more open, more available, and more usable. In chemistry, much data is produced by the workflows of chemistry research laboratories but the workflows are not overly focused on curating or publishing datasets [1]. Currently there is very little active support for research data management and deposit for broader access in chemistry [2]. Data sharing has not been a tradition in chemistry because the majority of research in chemistry falls under the "long-tail" of research [3]. Traditionally, in academia, individual chemistry labs often functioned as independent units and their research data were rarely shared, not even with their collaborators. In industry and corporate sectors, the proprietary data related to chemical substances were securely protected because of the potential value of the data, which could directly lead to products generating profit, such as new drugs and new materials. These traditions resulted in scattering of chemistry data in primary literature or locking them in proprietary screening databases.

However, the chemistry data published in primary literature were not completely unorganized. Because of chemists' needs for finding trustworthy data, indexing chemical data has been an essential task in curating chemical information. For example, Chemical Abstracts Service (CAS) started manually extracting and indexing data from primary literature in chemistry in 1907, transitioned to automated compound indexing in 1965 building from systematic nomenclature rules, and created an end-user interface to access the accumulative comprehensive databank in 1995, soon after computers enabled easy access and crosslinking of data [4, 5]. Earlier efforts were the independent Beilstein and Gmelin Handbooks, which

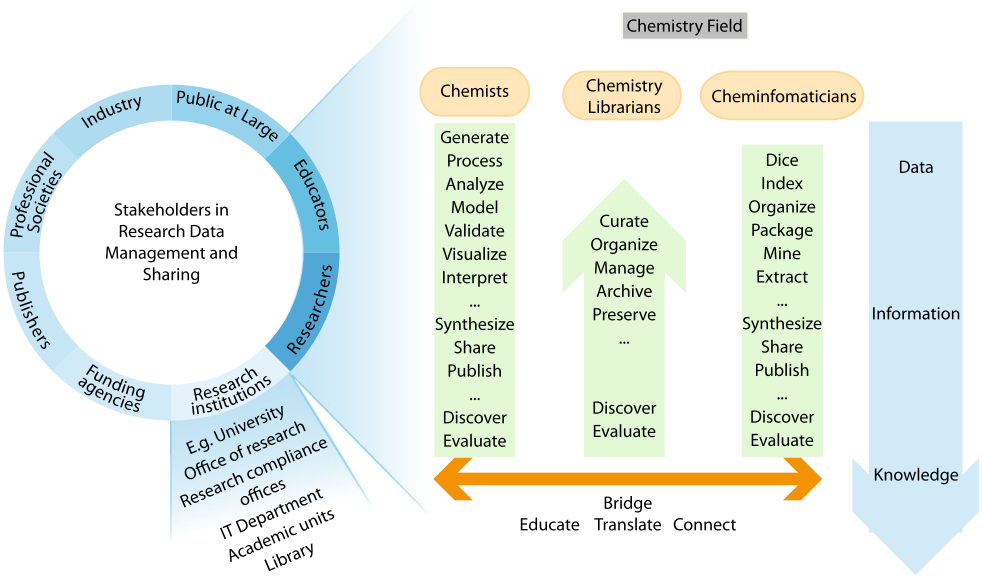

Fig. 1 Stakeholders in research data management [15, 16] and sharing and roles of chemists, cheminformaticians and chemistry librarians

started indexing chemical data and publication in the nineteenth century. Searching these indexed data and publications online by reactions and properties become available via the CrossFire system in 1994 [6], and the collective content can now be accessed through the Reaxys interface. Databases indexing chemical data and relevant publications, such as SciFinder and Reaxys, form a rich foundation of chemical data collection. However, these data are not readily available for automated data mining for several reasons: these proprietary databases provide very limited access for machine mining, there are not sufficient metadata and provenance associated with the data available to users since the targeted consumers have been humans not computers; and necessary metadata and provenance may not even exist in the primary literature that source the data in these compilations and indexes. In the face of open data and "big data" challenges, new infrastructure, workflows, practices, resources, and services need to be designed and implemented for the chemistry community. Although many collaborative initiatives to build infrastructures for different research purposes in chemistry are gradually being established by government agencies [7], professional societies [8], and private sectors [9], gaps still exist in supporting services for workflows and practices.

When funding agencies are asked for specific requirements on data management and sharing, they often direct the question to "be determined by the community of interest through the process of peer review and program management" [10]. Success in a particular "community of interest", such as the chemistry research community, lies in the buy-in of all the stakeholders to form a consensus over what and how to share research data. As shown in Fig. 1, we propose that "researchers and educators" in the academic chemistry environment refer to three sub-groups, namely chemists, chemistry librarians and cheminformaticians, each performing different but complementary actions over data and information. Librarians are applying expertise in curation, management, and preservation of documentation to data and gradually taking substantial roles in research data services [11, 12]. In the course of this work, chemistry librarians are becoming more involved with organizing, manipulating, and even mining chemical data, traditionally considered cheminformatics activities [13, 14].

Methodologies for engaging the chemistry community

Broad-scale chemical data capture requires more than just technical solutions. Academic chemists are overwhelmed with research pressures from funding to publication, quantity and quality of information already available, conflicting system interfaces and data file juggling. The current conversations around more and bigger data and community networks are not framed within these familiar challenges, and may even appear to compound them; re-use cases are not always readily apparent and thus are not considered high enough priority to compel change. Chemistry librarians working locally with chemistry researchers see that direct engagement within their experimental process is critical for successful data capture within and across individual lab groups. Chemists are struggling with lab and online information workflow and more granular and focused support is necessary to build up functional systems and services. Engaging a range of local support personnel including instrumentation facility managers, risk

management officers and teaching faculty as well as subject librarians is critical. Bridging engagement with data across established chemical information and literature processes beyond the lab and including other stakeholders such as scientific societies, publishers, other data compiling agencies, the chemical industry and the general public, is also an important factor for chemists. Chemistry librarians work with all of these groups in the course of their information support services (Fig. 1).

Chemistry librarians have always occupied a position at the interface between chemists and compiled data, published literature, and the systems used to organize, navigate and utilize these resources. It is not enough for quality information sources and systems to be available to trigger widespread and informed use. Just as reaction conditions are part of the chemical reaction, the human actions that put data into a computer database, get data out of it, make mistakes, remember or forget to use the system on either end, and the amount of time and training this takes are all part of the overall ecosystem of chemistry research. So are the librarians and other support personnel. In the course of addressing user questions and workflow challenges, librarians experiment with different ways of using these systems, often stumbling on inadvertent consequences of the indexing which they subsequently pass on to users in training. Librarians also assist users to formulate research problem statements to better inform the structures of their search questions based on the indexing of the databases [17]. These practices set the stage for librarians to consider the requirements for bridging chemists' experimental data workflow concerns with re-usable data structures.

In approaching the problem of transitioning chemical research workflows towards digital, we are interested in gleaning best practices that bridge common issues of data management with granularity of context and bring meaning to chemistry researchers and their immediate colleagues. We start this process with inquiry into the utility of existing information systems as used within various domains of chemistry research and practice. One method of surveying the landscape of data in a particular domain that has been developed by the library profession is to construct Data Curation Profiles (DCP) through interviews with researchers about the data lifecycle of a specific research project [18]. While the DCP approach has the potential to appreciate the full research background and identify the data curation needs of specific research projects, many DCPs need to be accumulated to appreciate the data landscape across a discipline. Data mining is another approach that derives patterns of data publication and presentation through analysis of research outputs.

We believe that a blend of the consultation and examination aspects of these approaches with additional review of the processes and practices of researchers and practitioners reveals a more holistic view of the data generation and flow within a domain and surfaces challenge areas that would especially benefit from digital data management intervention. In addition, we strive to profile discipline specific knowledge, assumptions, and practices so that other information specialists and computer scientists can build systems and artificial intelligence based on these explicit profiles. Here, we present two preliminary, independent explorations into this process examining the output, processes and perspectives of domain data producers and users. One exploration is examining reported data in chemists' publications to outline the general landscape of data in various sub-disciplines in chemistry. Another exploration is analyzing the landscape of information sources and uses within the domain of chemical health and safety to form the foundation of an information management system supporting safer chemistry research practice.

Understanding the landscape of chemistry data in publications

When we start to tackle the challenges of data curation, management, and preservation, one initial question we need to address is what are the kinds of data we need to curate, preserve, and help researchers to manage and share. As chemistry research becomes more and more interdisciplinary, defining what is chemistry and what kinds of data are considered chemistry data is not a straightforward question to answer. Traditional categorization of sub-disciplines such as organic, inorganic, analytical, physical and biological chemistry can no longer precisely describe research interests and characterize all the interdisciplinary effort. Instead, researchers tend to pursue one or more research themes such as energy science, optics and imaging, sensor science, RNA biochemistry, ultrafast dynamics, etc. [2].

To understand the landscape of data in a particular domain, information specialists are adopting two approaches. One is the DCP approach generating data profiles in structured documents. Since the approach was developed in 2007–2009, there have been 32 DCPs published in the DCP Directory across disciplines in humanities, sciences and social sciences [19]. More time is needed to accumulate sufficient DCPs to reveal the data landscape of any of these disciplines. Another approach is through mining data and texts of chemistry publications, such as the work done by Dr. Peter Murray-Rust's group [20]. This approach presents a more objective perspective by extracting the data related information from publications in chemistry and has the potential to run in large scale and outline the landscape of data in a particular discipline. In chemistry, the data and

information mining have been developed for both texts and graphics in publications, especially for chemical substance information, but manual curation is still needed to improve the automated mining results and the mining algorithms themselves [21].

For chemistry librarians, we have opportunities to adopt either the DCP approach or the data mining approach to understand our research communities but may be limited by time, resources, and technical skills to do so in isolation. However, we can leverage our immediate connection with both chemists and cheminformaticians to take advantage of both approaches and assist both parties with their goals in data management and sharing as well. Based on this idea, the Data Type Profiling (DTP) project [22] is designed to outline the landscape of chemistry data from a small but representative group of publications via a manual approach. The primary goal is to identify the various kinds of data reported by chemists in a variety of sub-disciplines of chemistry and the current practices in reporting these data. With the manually extracted results, we can describe the scope of data in chemistry to curate in a more explicit way with respect to different sub-disciplines. Reviewing these publications and data presenting practices with a critical eye will help us recognize potential issues in sharing these data in the future. In consultation with researchers, the statistics and noteworthy stories recorded by direct observations from the publications will form Data Type Profiles (DTPs) for each research lab and each sub-discipline of chemistry, which could further assist cheminformaticians to modify their algorithms for mining these data and their strategies for building infrastructure.

As a pilot study, we designed a workflow (Fig. 2a) and a Filemaker database (Fig. 2b) to manually compile DTPs for chemists in the Chemistry Department at the University of Michigan. Publications, primarily journal articles and book chapters authored by principal investigators (PIs) in the department, were retrieved from Web of Science and grouped by PI. Journal articles and associated supplemental information published during 2012–2013 were selected as the main reference set. Data and methods appearing in these publications were identified and described by filling in the attributes outlined in Fig. 2b. In order to describe the data types and methods in a consistent manner, the Chemical Methods Ontology (CMO) [23, 24] published by the Royal Society of Chemistry was employed. The ontology for Chemical Entities of Biological Interest (ChEBI) [25] was also used to describe the sample types consistently, but only at the high-level categories, such as organic small molecule, organometallic compound, inorganic molecular entity, polymer, peptide, protein, etc. In addition to recording observations of data types, methods, location, format, metadata, and provenance etc., those noteworthy practices of how particular types of data and methods were reported in publications were also recorded as additional notes. The manual description process made the best use of the capability of the human brain to summarize stories in presenting data and methods instead of focusing on details that machines might have compiled (false positives) or overlooked by fully automated mining of the publications.

One highlight of the database design is the separation of two entities, Data Reported and Experimental Method (Fig. 2b). When people discuss metadata standards for data, the impression might be that all pieces of metadata are associated with data. However, during the research lifecycle, data are also the output of research methods leading to interpretation, discussion, and conclusion, which also need description. The structure of CMO reflects the same consideration in terms of the relationship between data and methods. Often metadata and provenance describing experimental conditions, instrument types and parameters, date and time, analysis process, are more naturally associated with the research methodology used instead of a particular data instance, especially when the data reported have been much processed from initial collection. In addition, clarifying the relationship between the data and methods reported helps identify the stage of data during its lifecycle. For example, the diastereomeric ratios can be calculated from output of multiple types of measurement methods including NMR of the diastereomeric mixture, chromatographic separation, chemical decomposition, and comparison of optical property to known substance. Depending on the methodology used to obtain the diastereomeric ratios, the metadata and provenance to be recorded would be different for the same piece of data. Therefore, it is necessary to analyze both data and method types often presented in a particular sub-discipline so that the resulting DTPs can be used for establishing best practices and metadata standards.

As preliminary results, the word cloud plots in Fig. 2c partially illustrate a DTP of an organometallic chemistry lab based on five publications from this lab during 2012–2013. The relative sizes of the texts reflect the relative frequencies of different types of data and methods presented in the publications and their formats as well as locations. As shown in the figure for this lab, the types of data reported most often were substance structures, reaction schemes, ^{1}H-NMR, ^{13}C-NMR, yields, diastereomeric ratios, mass spectrum etc. The majority of the data were only available in PDF format with only one occurrence out of 52 being a reusable format—CIF (crystallographic data). About half of the data occurrences were found in the supplemental material. The locations of those data were often found in the order of texts, figures, tables and captions while one out of 52 occurrences was linked out to an external data repository via a unique identification number in the repository,

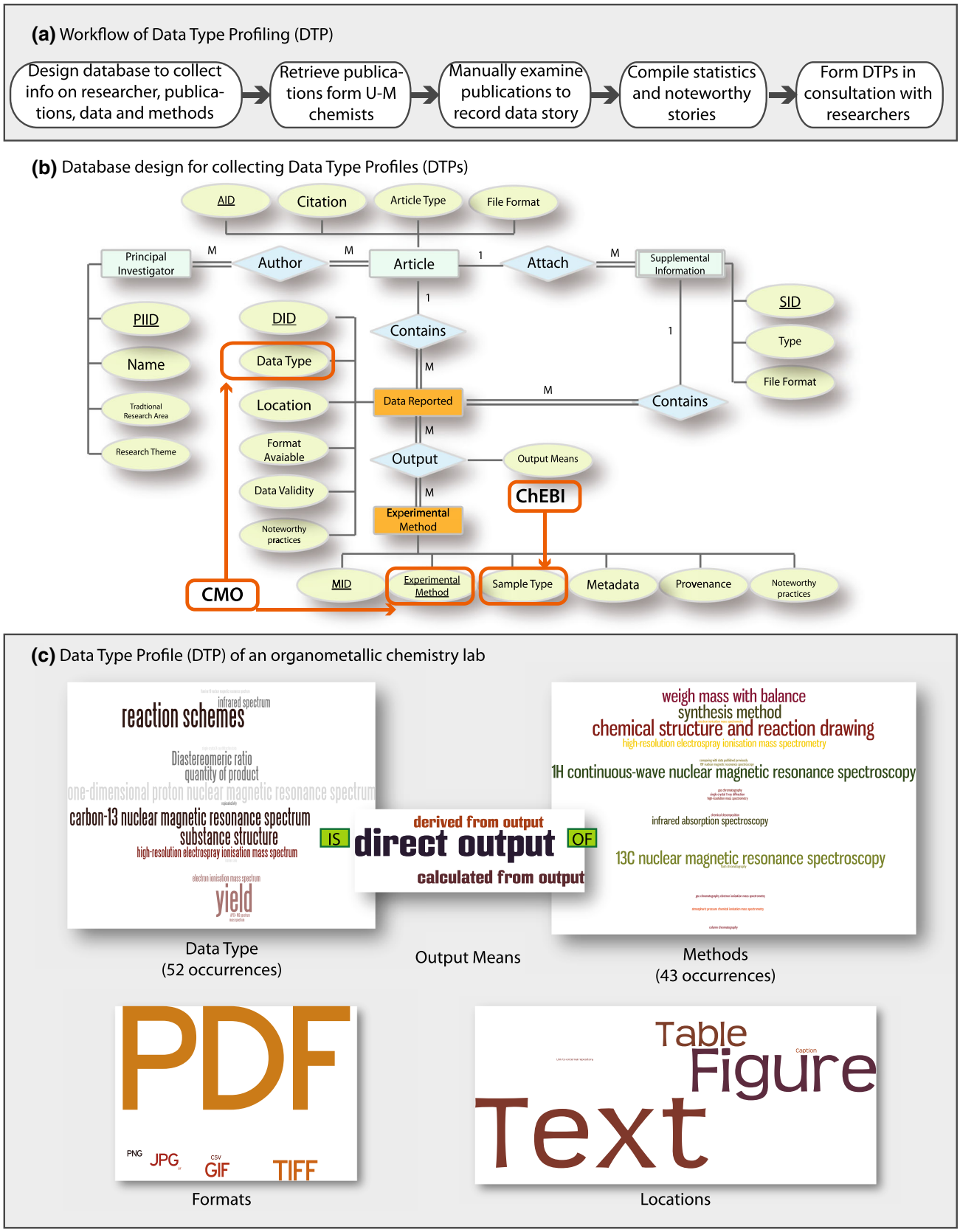

Fig. 2 Workflow (**a**), database design (**b**), and some preliminary results (**c**) of the Data Type Profiling (DTP) project

Cambridge Crystallographic Data Center (CCDC). No articulation of accuracy and precision for any pieces of data were presented. Among the 43 occurrences of methods mentioned in the publications, less than half of them had some equipment information and some experimental plan specific information reported, such as solvent used. Mostly, there were no date/time stamps, no replications, no data analysis processes described for the methods. This overall picture of an organometallic chemistry lab suggests the potential of using these observations for metadata standard development and data management education design.

Meanwhile, the manual process also allows us to provide crucial feedback to cheminformaticians on their foundational work. Since CMO was used extensively when we recorded the data types and methods, we were able to record the data types and methods not yet indexed in the ontology, synonyms of data types of methods not appearing in the ontology, and those implicit relationships between data output and methods. All of the above can be used to enrich the ontology.

We also recorded some noteworthy data reporting practices, which can be useful for cheminformaticians especially doing mining. For example, for developers of spectra recognition algorithms, it is helpful to know that ^{1}H and ^{13}C NMR data were often presented as both chemical shifts and with the spectra graph attached in supplemental materials but ^{19}F NMR data were only reported as chemical shifts. This practice may have resulted from journal publishers' requirements of supplemental material, which may warrant adjustment to suit emerging needs in data science. There are other noteworthy practices potentially useful in improving data mining. For instance, yields were only reported in the figures and tables, which were mostly not machine interpretable images. Specific quantities of products only accompanied the corresponding yields in the supplemental information but not in the article. When a group of the same type of data is being presented together in an organized manner, they may reveal another "type of data", e.g. a table listing yields of reactions with different ligands in the reactants could demonstrate ligand effects. Sorting out this kind of data types will augment mining strategies.

The DTP described above was from only one particular organometallic chemistry lab. We are in the process of compiling DTPs for more labs in various sub-disciplines[1] and research themes[2] in the Chemistry Department of the University of Michigan. By completing these DTPs, we will obtain a representative scope of the common types of data to the interest of chemists in various sub-disciplines. Next, comparing these DTPs with DTPs of chemists from other research institutions generated from the same profiling protocols will form a comprehensive data landscape in chemistry and the foundation to define the scopes, standards, and best practices as well as help establish infrastructure for data sharing in chemistry.

This manual approach to examining the practice of data and research results presentation can be applied to other projects relevant to data and information mining, management and sharing. One example is the ongoing collaborative project, ChemReader, at the University of Michigan [26], where the manual examination of presentation of chemical structures in biomedical literature could provide clues to improve the image-to-structure algorithm and further development of artificial intelligence. Another example is identifying information needed in supporting safe chemistry research practices as described in the next section. Through these projects, chemistry librarians are truly acting on our bridging roles and developing long-term engagement practices in the data management and sharing process.

Managing chemical data to support a use case in chemical risk planning

In addition to mapping the landscape of published data in chemistry, librarians are also studying data management and digital workflows. One opportunity is to consider a case for re-using research data to support compelling needs such as planning for greener and safer chemistry practice. Managing information to support chemical risk management is primarily associated with chemical labeling at this point in time, and some preliminary data management work has begun in this area through the Global Harmonized System of Classification and Labelling of Chemicals (GHS) [27, 28]. However, broader evaluation of hazard implications under various conditions, management of potential exposure and risk through protective equipment and molecular substitution, and incident analysis to inform best laboratory practices are also relevant for experimental planning. Just as chemical structure and activity data are used to streamline the process of discovering viable drug candidates, approaching the problem of managing chemical risk quantitatively through chemical reactivity and process analysis is suggestive of cheminformatics application. Enabling incorporation of hazard reactivity and risk management data into the research workflow at the level of experimental design can promote a culture of safety among

[1] The sub-disciplines and faculty members working on these sub-disciplines at the Department of Chemistry, University of Michigan are listed on http://www.lsa.umich.edu/chem/people/faculty/facultybyresearchcluster (Accessed April 2014).

[2] The faculty members and their self-reported research themes at the Department of Chemistry, University of Michigan are listed on http://www.lsa.umich.edu/chem/people/faculty/facultybyresearchthemes (Accessed April 2014).

those closest to the process and the outcomes, researchers and students [29].

Following several high profile laboratory events, a report of the US Chemical Safety Board on safety in academic chemistry laboratories recognized a need for an organized risk assessment process that meets the needs of the research laboratory [30]. A significant challenge to implementing risk assessment strategies in the decentralized academic research setting is the effort required to collect and organize the disparate information necessary and detail how it applies to a specific chemical process or laboratory. The American Chemical Society (ACS) Committee on Chemical Safety reviewed the Recognize, Assess, Manage, Plan (RAMP) chemical risk management model originally developed for education in the undergraduate class setting [31, 32], and the ACS Division of Chemical Health and Safety (CHAS) is partnering with the Division of Chemical Information (CINF) to assess the information available for supporting use of this model in the academic research environment. The primary goal is to formulate requirements for information organizational structures such as shared vocabularies and descriptions to facilitate discovery of disparate information and the management of further data and documentation produced by the laboratory safety profession in support of broader dissemination of best practices [33].

There are many approaches employed in industrial settings for incorporating risk managment into process design, including hazard and operability studies (HAZOP), and management of change (MOC). Chemical, laboratory and facility level processes are thoroughly and repeatedly tested as they are scaled, optimized and validated. While chemical processes are not standardized in research settings, there are standard operating procedures for known hazards, and there is certainly opportunity to incorporate hazard analysis into experimental planning based on documented reactivity and incident reports. For example, using specific handling and transferring procedures for pyrophoric reagents such as *tert*-butyllithium and other compounds known to be reactive at standard laboratory atmosphere and pressure, and subsequently adjusting for planned alterations and appropriately preparing against unplanned alterations in these conditions [34]. However, associating hazards and operating procedures with individual compounds only is not adequate; material incompatibility and ripple effects from process change also contribute to the overall hazard level of an experiment [35]. A poignant example is azide chemistry; not only are azide reagents incompatible with various metals and chlorinated solvents, highly toxic and shock sensitive hydrazoic acid can be readily generated and explode. Traditionally this type of experiment is managed through engineering controls such as fume hoods and glove boxes, but can still result in breach of safety even at academic scales [36]. It is also possible to quantitatively manage risk up front by minimizing generation of hazardous compounds and conditions through analysis and adjustment of molecular design, process operations and conditions [37].

These approaches require experimental literature review and experimental data management throughout the process planning and execution. To document an information ecosystem of chemical process planning in the context of risk management, we framed an information flow through the RAMP risk assessment process based on the model of identifying and using relevant information in research as outlined by the American College and Research Libraries Information Literacy Competencies [38]. Researchers would actively engage in determining the scope of information needed, accessing it efficiently, critically evaluating the sources and content, and incorporating it into their experimental design and overall risk management practice. The process is interactive and iterative over ongoing experimental and laboratory processes (represented by an "i" in iRAMP). Figure 3 illustrates the information ecosystem surrounding an iRAMP process flow (adapted) [39]:

1. chemical process description
2. data collection and presentation
3. chemical safety level (CSL) assessment and risk management determination
4. documentation of process and supporting data
5. collection into hazard assessment process warehouse

In addition to chemical process description as currently recorded in disparate laboratory notebooks, relevant planning data can source from peer reviewed literature including articles and data compilations covering novel compounds, transformations, condition optimization, experimental methods, standard protocols and cautionary notations; official reports including annotated reviews, data compilations, hazard and toxicological classification systems, and incident reports; chemical supplier documentation, including safety data sheets (SDSs); and local environmental health and safety data including job hazard analyses, chemical inventories, inspection checklists, and environmental monitoring [40–43]. Parsing, linking and reframing of diverse and disparate source data into the immediate context of an individual experiment for the purpose of informing risk assessment requires appropriately controlled tagging [44]. We are investigating the potential of existing and possible new chemical ontologies to support data infrastructure for risk management applications across the global chemical space [23–25, 28]. Many biological and physical hazards associated with chemical compounds are found in ChEBI; CMO might be expanded to include laboratory operation and procedure concepts involved in process planning; types of chemical and

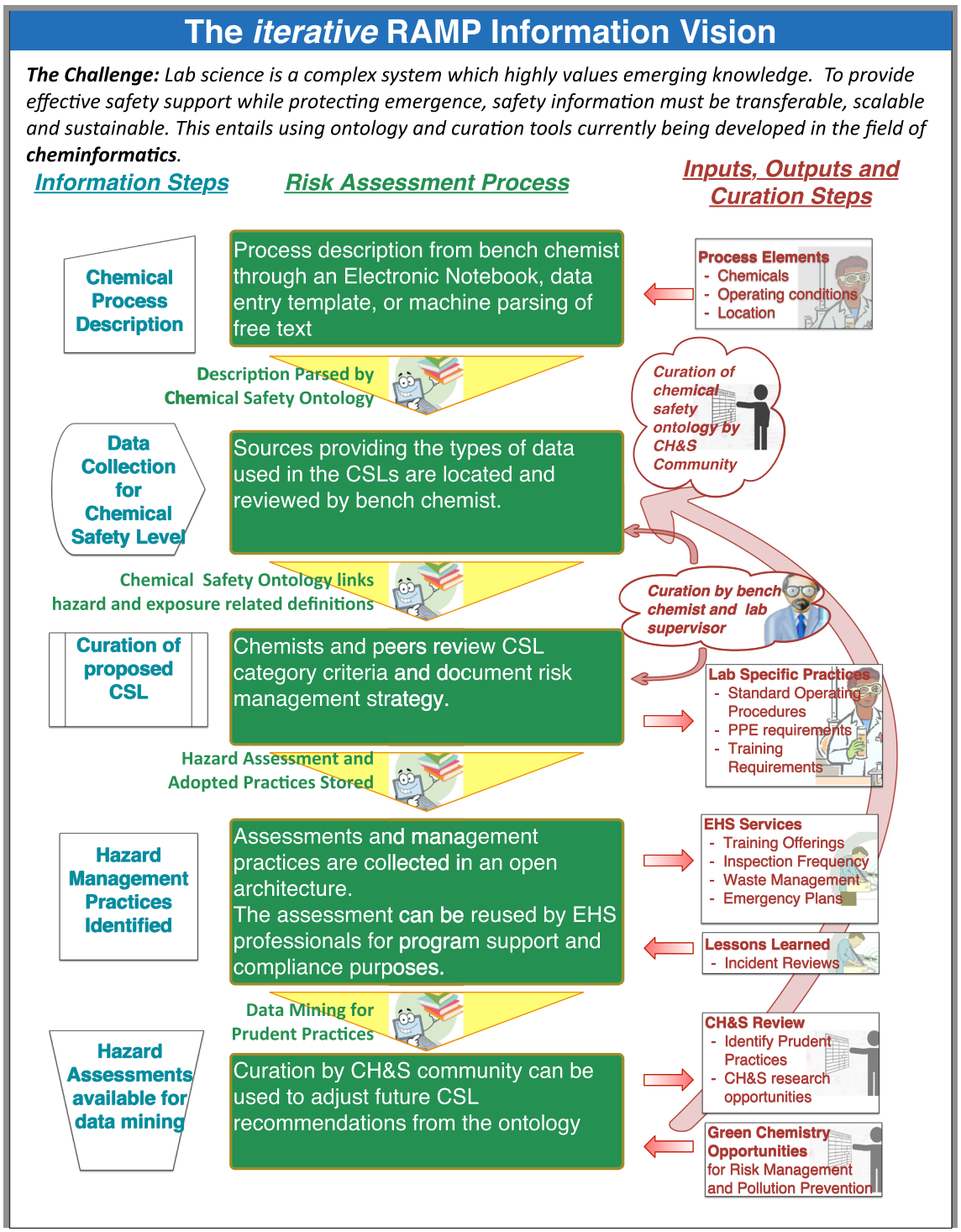

Fig. 3 iRAMP process information workflow

laboratory events for classifying lessons learned and mitigation responses might be captured in yet another ontology.

Chemical safety oversight involves chemical process specific risk assessment at the molecular and chemical methodological levels, through synthesis planning using safer and greener reagents, and at the local laboratory and facility levels through storage, handling, waste management and personal protection. To facilitate this process, the ACS Chemical Safety report considers the potential of an assessment rubric based on the control banding approach

Table 1 Chemical safety levels (CSL, preliminary partial view)

CSL	Conceptual hazard description	Chemicals used	Explosives	Fire	Oxidizers
1	Chemical hazards equivalent to typical household kitchen	Consumer products in consumer packaging or equivalent	Not appropriate for use at this CSL	Not appropriate for use at this CSL	Not appropriate for use at this CSL
2	Chemical hazards equivalent to teaching lab settings (restricted chemical inventory; well-established, supervised procedures in place)	Low concentration acids/bases, simple alcohols, solid salts, simple asphyxiant compressed gases	Not appropriate for use at this CSL	Warning	"Warning" liquids and solids
3	Varying laboratory hazards within an identified range (open hazardous chemical inventory; evolving procedures)	Flammable solvents, corrosives, toxics, flammable gases. No air/water reactive, pyrophoric materials	Warning	Danger	"Danger" liquids and solids
4	Novel hazards or severe established hazards (high hazard chemicals or processes with well established procedures)	Air/water reactive, pyrophoric materials or gases. Explosives or potentially explosive compounds, highly toxic materials	Danger	Flammable gases	Oxidizing gases

used in industrial laboratory settings [31, 45]. Each CSL describes conceptual hazard scenarios and associated types of chemicals, ranging from a typical household kitchen with consumer products up through situations involving novel or established severe hazards such as pyrophoric materials (see partial preliminary version in Table 1). Hazard classification terminology and description are based on the GHS and SDS specifications. Subsequent risk management strategies are associated with safety levels, with additional dependencies for more specific scenarios, such as fire-resistant lab coats and elevated ventilation requirements.

Capture of experimental level CSL and risk management determinations, supporting data and any subsequent hazard related observations in the course of the experiment with full provenance metadata could enable information management for campus level safety programs including standard operating procedures, training requirements, personal protective equipment recommendations, emergency plans, and incident reporting. Shareable data deposited in a hazard assessment warehouse could support further data mining by chemical researchers and EHS professionals and maximize benefit from lessons learned for all stakeholders. Accumulated data would also support further curation of the iRAMP process flow, CSL levels, the underlying ontologies, and other safety research and practice recommendations by the Chemical Health and Safety community at national and international levels [46].

The scope of this effort is aimed at the chemistry community level, focusing on bridging opportunities for technical applications with domain information flow and diverse human user groups and requirements. Analysis of the chemical health and safety information landscape is an ongoing endeavor for supporting establishment of prudent practices in experimental process planning [47]. The number of information streams, the variety of output requirements, the diversity of stakeholders and the curation considerations for data sharing suggests the need for community engagement. We are looking at other community curation models such as that supporting ChEBI content with an eye towards building a distributed network of contributors, curation tools and documentation, presentation and education materials, and a community development repository to support the local endpoint needs of the stakeholders. While the project thus far has focused on the scenario in US laboratories, similar concerns are being considered in Europe and elsewhere and we are investigating opportunities to extend collaborations. The initial development of an information management roadmap and meta-structures is a happy confluence of the needs to transition current and traditional information sources into the digital environment and the essential requirement to

engage both chemical safety and information professionals; a logical extension of the current professional service capacity and stewardship foci of both communities.

Challenges for chemistry librarians

Through our independent explorations, we see four related but distinct challenges surfacing that suggest potential areas of involvement for chemistry librarians in data management and sharing: engaging chemists, curating chemistry metadata, process curation and connecting stakeholders.

From the human point of view, chemistry has a long history of intra-discipline developed data driven research approaches based on existing data enriched systems (e.g. SciFinder, Reaxys, and other proprietary screening databases), which incorporate fairly extensive documentation. These practices are inherent to the chemistry culture, not alien. Some data organization and interpretation practices, such as graphical representation, are so inherent that chemists may experience initial difficulty making more of the process explicit to meet digital representation requirements [48]. However, chemists do appreciate good data handling when they see it, and are sensitive to well captured and re-presented nuance and interpretation when it impacts the science, such as that produced, compiled and curated by the CCDC and the National Institutes of Standards and Technology [49, 50]. There is already some precedent for capturing other characterization data at the point of manuscript publication as supplementary information as well as emerging best practice recommendations for handling this information for publishers [51]. We would argue that the concept of accompanying data is not alien to chemists and the potential for increasing engagement in digital deposit is there. The problems are well understood and documented on technical fronts [52]. What is needed for chemists are better support scenarios across the value chain of research and publication.

Robust capture, representation and further manipulation of meaningful data require robust description and organization. It would appear that there is very little discipline wide compiled metadata or other useful organizational structure in readily useful and available form that pertains to chemistry. What is available in digitally enabled structures is mostly general, at the bibliographic or data element level, much less at the chemical reactivity or process levels that are the most familiar and of direct interest to research chemists. Chemical reactivity represents their science, chemical processes represent their workflow. Description that supports analysis and inference of chemical reactivity is of interest to their research agenda, and has been a primary focus of the cheminformatics field. Description that supports analysis and inference of chemical processes that can improve lab workflow as well as safer and greener experiments is a more recent consideration. However, chemists have long been interested in these questions, and even before the advent of computers to help with scale, they were devising systematic description, classification and flowcharts. These are still captured in the long-standing work of standards [53] and government organizations [54], as well as various other data and literature compilations, some captured in XML awaiting parsing, some in PDF awaiting improved mining, and much more still sitting in libraries in hardcopy. We suggest that further review of what structures are in use by chemists and other chemistry professionals can lend insight into improving both retrospective mining and new data systems design.

The act of engaging in data capture using appropriate metadata is part of the curation process, currently conducted in disparate fashion in hardcopy notebooks, electronic data sheets, directly through instrument feeds, etc. Further checking of scientific merit, experimental accuracy, file management and copyediting are also part of the curation process. These activities are already happening in all research labs in a variety of ways involving a host of supportive infrastructures, including professional personnel. These positions in research academies are detailed to overseeing high-level functionality of various processes critical to research, including instrumentation, environmental health and safety, training and literature searching. Digital capture and management of chemical data will necessarily involve all of these processes and personnel; it is a natural extension of their professional work. Much of the attention toward creating appropriate digital data cultures that both engage their chemistry colleagues and ensure consistent data management could be addressed at this level with appropriate tools and training, such as modeled by the ChEBI ontology project [25].

Curation of data continues well beyond the lab and host research institutions as data are published in various forms and venues, further compiled and checked, and sourced into a variety of applications such as diagnostic tools, teaching materials, and engineering specifications. Published data at many of these points are also captured by libraries for direct access by their current user communities, and archived for future communities. Traditionally many of these parties handling data in some form have been siloed, acting independently at points along the data cycle that suit their immediate purpose. Workflows are likewise optimized to immediate purpose and in isolation of impacts that might ripple through the cycle, such as mandates for deposit. Some coordination does occur within particular agreements, such as between data sources and compilers like Acta Crystallograhica and the CCDC, where the rationale can be closely tied to the immediate value

proposition. These cases might serve as models both for sharing best practices of data management, and also how such cooperative approaches might play out in other well-scoped scenarios, such as pipelining deposit of other types of characterization data.

The prospect of open data might seemingly bypass many stakeholders currently involved in the data cycle. On a pragmatic level the state of the established data flow guides the course of research; enabling increased engagement in useful directions by as many stakeholders as possible has the potential not only to channel more data into useful places and in useful forms, but also to reach chemists at critical points in their current workflows impacted by these parties. There are enough pain points in the established overall flow experienced by all parties to suggest many opportunities for win–win adjustment scenarios. As stakeholders broker more connections in multiple directions in their own workflows, such positive effects have better potential to ripple throughout the system, and challenges would be more likely collectively addressed instead of collectively ignored, as has been the case with the majority of supplementary information management.

The value of librarians lies in an overarching view of the data life cycle garnered in the course of supporting the implementation and use of a wide range of systems available in the research world of practicing chemists. Librarians are stakeholder-agnostic, cross-perspective and do not specialize overly in any one area. We rely heavily on the complementary expertise of other stakeholders in our workflows. We are most interested in collaborative solutions to transitioning chemical research workflows towards digital. We envision a multi-pronged approach to this scenario: liaising with systems developers and individual lab groups who want to make their particular processes robustly digital, bringing data management into regular research practice and education, and contributing to information organization motifs. We offer a *pragmatic* stewardship view of the research data management cycle, what are the actual states of the data, and what issues need to be addressed at each stage and throughout the cycle [55]. Pioneering scientists who are willing to pilot management systems and tools on an experimental level can provide case study environments and serve as ambassadors to colleagues and other stakeholders such as publishers.

Towards digital research workflows for chemists

Beyond data management and sharing, librarians aspire to provide resources and services for chemists throughout the whole research lifecycle (RLC) [55, 56]. To address the four challenge areas for chemistry librarians outlined above, we need to broaden our perspective of curation across the whole research workflow instead of focusing just on the end products—datasets and publications. Only if we understand workflows on a holistic level can we then identify how research data, records, and output flow though the RLC and determine the best entry points to engage chemists in curating the process. An ideal information management environment serving the workflow of a chemist would enable them to actively participate and allow chemistry librarians and other information specialists to give input during the process. The management approach that chemists need now is beyond an electronic lab notebook in the traditional sense. The ideal system should be a modular suite of tools to add, organize, manage, index, search, archive, export, and share a variety of different objects at any given time and spatial points. These objects can appear in any format and at any location during the research process. You may or may not be clear about the destination and purpose of an object when it appears; and each object and the associated metadata can be repurposed for reuse or new, unanticipated uses. Recording appropriate metadata and provenance for all of these objects seems impossible to be planned, although worthy efforts are underway [57].

Some computer scientists have analyzed the scientific workflow based on business workflow models and attempted to establish models to streamline the management and sharing of data and their provenance, especially for those fields with comparatively homogenous workflows across different labs [58]. Since chemistry is such a diverse field and intersects with so many other disciplines, research workflows in chemistry tend to be very complex and heterogeneous. Based on our interactions with chemists, the ideal workflows are different from lab to lab, or even from individual to individual. Academics value the diversity of scientific approaches and consider it crucial to excite innovative ideas and breakthroughs. Therefore, we are not attempting to standardize human workflows. Instead, we use the RLC as a model to identify those data, records, outputs and other information that are important to each step of the RLC, and their possible formats and locations (Fig. 4). This model is a starting point for us to translate the needs of chemists and provenance of data into tangible objects and processes to curate and manage at various stages of the RLC. It is also the first step by which we can connect chemists and computer scientists to establish a flexible system for chemists to manage the whole research workflow intuitively. What's more, the model can be used for identifying specific areas where librarians could provide consultation, instruction, and education services for researchers.

Librarians are extending beyond our traditional comfortable zone of curating the "packaged" information, such as books and articles, and moving into curation of a variety

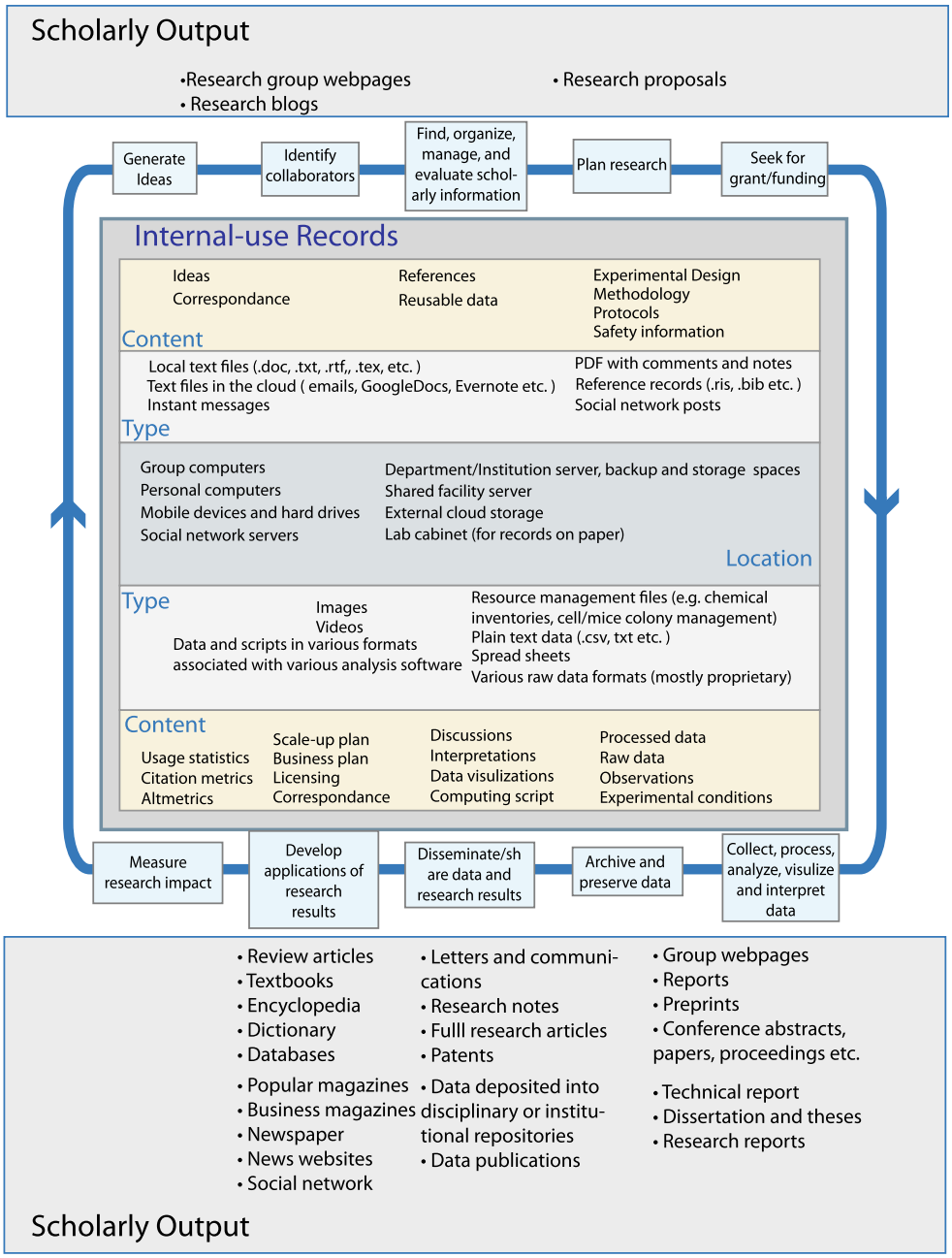

Fig. 4 Internal-use records and scholarly output from researchers in chemistry throughout the research lifecycle

of research records and output as well as the overall research process itself. We can choose where to start based on the analysis of the research workflow in Fig. 4 and the chemistry research environment of our local institution. At the micro level, we can choose specific objects to study and develop better understanding of scope, metadata needed, provenance to be recorded, etc. The DTP project we discussed is a good example of choosing data presented in publications as one type of object for further investigation. At the macro level, we can start to compile various types of workflows involving different RLC steps, objects, and flows of objects within the context of one type of information problem. These workflow scenarios can become part of the foundation to design flexible systems for research management. The chemical risk management use-case we presented demonstrates this possibility and our potential for initiating and continuously participating in these multi-faceted ideas. If our peer librarians, cheminformaticians, chemists, and computer scientists all choose different objects and processes in the workflow and work together from different perspectives, we can be reasonably optimistic that we will build a flexible collaborative network for research management in the near future [59]. During this process, chemistry librarians will also continue

developing data literacy, research literacy and support programs to ensure that research practices and culture in chemistry communities grow healthier.

In their recent review article, Bird and Frey [52] comprehensively analyzed how chemistry, cheminformatics, computer science, and information technology could work hand-in-hand to enhance reproducibility, sharing and collaboration in chemistry in the new era of open science and e-Research. In our opinion, librarians have an active role in bridging, connecting, and translating among chemists, cheminformaticians, and computer scientists. We are and will continue to apply and challenge our professional skills in all of these areas in our pursuit to improve the scientific research and learning environment. Chemistry librarians can be selectors, organizers, curators, archivists, managers, coordinators, "translators", analysts, trainers, and also educators in the process of serving the research lifecycle. We are also researchers ourselves navigating through the scientific universe. We continue developing our core competences in managing data and information as well as making connections among people, resources and stakeholders with the goal of contributing to a more open, efficient, healthy and collaborative research environment.

Acknowledgments The authors would like to thank colleagues at several institutions including the Cornell University and University of Michigan Libraries and Departments of Chemistry respectively, the American Chemical Society, and the Royal Society of Chemistry; particularly Ralph Stuart in the Environmental Health and Safety Department at Cornell University for collaborating on the iRAMP project; and Dr. Kazuhiro Saitou, Dr. Gus Rosania, and Dr. Jungkap Park at the University of Michigan for the collaboration opportunity on the ChemReader project.

References

1. Carol T, Suzie A, Kimberly D, Aydinoglu AU, Wu L, Read E, Manoff M, Mike F (2011) Data sharing by scientists: practices and perceptions. PLoS One 6(6). doi:10.1371/journal.pone.0021101
2. Li Y, Tschirhart L (2012) Preparing To support research data sharing. In: Xiao N, McEwen LR (eds) Special issues in data management, vol 1110. ACS symposium series. American Chemical Society, pp 145–162. doi:10.1021/bk-2012-1110.ch009
3. Velden T, Lagoze C (2009) Communicating chemistry. Nat Chem 1(9):673–678. doi:10.1038/nchem.448
4. Weisgerber DW (1997) Chemical abstracts service chemical registry system: history, scope, and impacts. J Am Soc Inf Sci 48(4):349–360. doi:10.1002/(sici)1097-4571(199704)48:4<349:aid-asi8>3.0.co;2-w
5. Ridley DD (2009) Front matter. In: Information retrieval: SciFinder®. Wiley, pp i–xii. doi:10.1002/9780470749418.fmatter
6. Meehan P, Schofield H (2001) CrossFire: a structural revolution for chemists. Online Inf Rev 25(4):241–249. doi:10.1108/14684520110403768
7. PubChem. http://pubchem.ncbi.nlm.nih.gov/. (Accessed Apr 2014)
8. ChemSpider. Royal Society of Chemistry. http://www.chemspider.com/. (Accessed Apr 2014)
9. Open PHACTS: Open pharmaceutical space open PHACTS consortium. http://www.openphacts.org/. (Accessed Apr 2014)
10. Data Management and Sharing Frequently Asked Questions (FAQs). National Science Foundation. http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp. (Accessed Apr 2014)
11. Tenopir C, Sandusky RJ, Allard S, Birch B (2013) Academic librarians and research data services: preparation and attitudes. IFLA J 39(1):70–78. doi:10.1177/0340035212473089
12. Environmental Scan 2013 (2013) Association of College and Research Library, Research Planning and Review Committee. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/EnvironmentalScan13.pdf. (Accessed Apr 2014)
13. Willett P (2008) From chemical documentation to chemoinformatics: 50 years of chemical information science. J Inf Sci 34(4):477–499. doi:10.1177/0165551507084631
14. Warr W (2011) Some trends in chem(o)informatics. In: Bajorath J (ed) Chemoinformatics and computational chemical biology. Methods in molecular biology, vol 672. Humana Press, New York, pp 1–37. doi:10.1007/978-1-60761-839-3_1
15. William KM, Suzie A, Amber EB, Robert C, Kimberly D, Mike F, Steve K, Rebecca JK, Carol T, David AV (2012) Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. Ecol Inform 11:5–15. doi:10.1016/j.ecoinf.2011.08.007
16. Erway R (2013) Starting the Conversation: university-wide research data management policy. OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013-08.pdf. (Accessed Apr 2014)
17. Currano JN, Roth D (eds) (2014) Chemical information for chemists: a primer. Royal Society of Chemistry, Cambridge
18. Data Curation Profile Toolkit. Purdue University & University of Illinois at Urbana-Champaign. http://datacurationprofiles.org/. (Accessed Apr 2014)
19. Data Curation Profiles Directory. Purdue University. http://docs.lib.purdue.edu/dcp/. (Accessed Apr 2014)
20. Townsend JA, Adams SE, Waudby CA, de Souza VK, Goodman JM, Murray-Rust P (2004) Chemical documents: machine understanding and automated information extraction. Org Biomol Chem 2(22):3294–3300. doi:10.1039/b411033a
21. Gurulingappa H, Mudi A, Toldo L, Hofmann-Apitius M, Bhate J (2013) Challenges in mining the literature for chemical information. RSC Adv. doi:10.1039/c3ra40787j
22. Li Y (2014) Profiling common types of research data produced by chemists at the University of Michigan. In: 247th ACS national meeting and exposition, American Chemical Society, Dallas, TX, USA
23. Batchelor C (2014) Chem Methods Ontol http://purl.bioontology.org/ontology/CHMO. (Accessed Feb 2014)
24. Chemical Methods Ontology (CMO) R Soc Chem. http://www.rsc.org/ontologies/CMO/. (Accessed Feb 2014)
25. The database and ontology of chemical entities of biological interest. European Bioinformatics Institute, European Molecular Biology Laboratory. http://www.ebi.ac.uk/chebi/. (Accessed Apr 2014)
26. Park J, Rosania GR, Saitou K (2009) Tunable machine vision-based strategy for automated annotation of chemical databases. J Chem Inf Model 49(8):1993–2001. doi:10.1021/ci900029v
27. Globally harmonized system of classification and labelling of chemicals (GHS) (rev. 5) (2013) United Nations Economic Commission for Europe. http://www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html. (Accessed Apr 2014)
28. Borkum M Machine-processable representation and application of the Globally Harmonized System. In: 247th ACS national

meeting and exposition, American Chemical Society, Dallas, TX, USA
29. Kemsley J (2014) Chemistry professors promote lab safety. Chem Eng News 92(23):30–31
30. Experimenting with Danger. U.S. Chemical Safety Board. http://www.csb.gov/videos/experimenting-with-danger/. (Accessed Apr 2014)
31. Identifying and evaluating hazards in research laboratories (2013) American Chemical Society Committee on Chemical Safety. http://www.acs.org/content/dam/acsorg/about/governance/committees/chemicalsafety/identifying-and-evaluating-hazards-in-research-laboratories-draft.pdf. (Accessed Apr 2014)
32. Hill RH, Finster DC (2010) Laboratory safety for chemistry students. Wiley, Hoboken
33. The iRAMP Development Blog. http://www.irampp.org/blog/. (Accessed June 2014)
34. Busacca CA, Eriksson MC, Haddad N, Han ZS, Lorenz JC, Qu B, Zeng X, Senanayake CH (2013) Practical synthesis of di-tert-butyl-phosphinoferrocene. Org Synth 90:316–326. doi:10.15227/orgsyn.090.0316
35. Urben PG (ed) (2007) Bretherick's handbook of reactive chemical hazards. Elsevier, Amsterdam
36. Kemsley J (2014) Explosion injures University of Minnesota graduate student. American Chemical Society. http://cenblog.org/the-safety-zone/2014/06/explosion-injures-university-of-minnesota-graduate-student/. (Accessed June 2014)
37. González-Bobes F, Kopp N, Li L, Deerberg J, Sharma P, Leung S, Davies M, Bush J, Hamm J, Hrytsak M (2012) Scale-up of azide chemistry: a case study. Org Process Res Dev 16(12):2051–2057. doi:10.1021/op3002646
38. Information Literacy Competency Standards for Higher Education. Association of College and Research Libraries, American Library Association. http://www.ala.org/acrl/standards/informationliteracycompetency. (Accessed June 2014)
39. Safety in Research Laboratories_UC CLS Workshop 2014 UC Center for Laboratory Safety. http://cls.ucla.edu/2013-12-07-09-26-14/cls-workshop-2014. (Accessed June 2014)
40. Wrublewski D (2014) Lab safety—chemistry—libguides at Caltech. http://libguides.caltech.edu/content.php?pid=58674&sid=3626650. (Accessed Apr 2014)
41. Baysinger G (2014) Lab safety—guides—Stanford University Libraries. http://library.stanford.edu/guides/lab-safety. (Accessed Apr 2014)
42. Conneting Chemistry and Safety. American Chemical Society, Division of Chemical Health and Safety. http://www.dchas.org/. (Accessed Apr 2014)
43. Lab and Research Safety. Cornell University, Environmental Health and Safety. http://sp.ehs.cornell.edu/lab-research-safety/Pages/default.aspx. (Accessed Apr 2014)
44. Stuart R, Toreki R (2014) Learning opportunities in three years of hazmat headlines. J Chem Health Saf 21(2):2–8. doi:10.1016/j.jchas.2013.11.002
45. Bassan E, Ruck RT, Dienemann E, Emerson KM, Humphrey GR, Raheem IT, Tschaen DM, Vickery TP, Wood HB, Yasuda N (2013) Merck's reaction review policy: an exercise in process safety. Org Process Res Dev 17(12):1611–1616. doi:10.1021/op4002033
46. UC Center for Laboratory Safety. https://cls.ucla.edu/. (Accessed June 2014)
47. Board on Chemical Sciences and Technology, Division on Earth and Life Studies, National Research Council of the National Academies (2011) Prudent practices in the laboratory: handling and management of chemical hazards, updated version. The National Academies Press, Washington, DC
48. Brecher J (2008) Graphical representation standards for chemical structure diagrams (IUPAC recommendations 2008). Pure Appl Chem 80(2):277–410. doi:10.1351/pac200880020277
49. NIST Standard Reference Data National Institute of Standards and Technology. http://www.nist.gov/srd/. (Accessed Apr 2014)
50. Cambridge Crystallographic Data Center. http://www.ccdc.cam.ac.uk/pages/Home.aspx. (Accessed Apr 2014)
51. David PM (2012) Supplemental journal article materials. In: Special issues in data management, vol 1110. ACS symposium series. American Chemical Society, pp 31–45. doi:10.1021/bk-2012-1110.ch003
52. Bird CL, Frey JG (2013) Chemical information matters: an e-research perspective on information and data sharing in the chemical sciences. Chem Soc Rev 42(16):6754–6776. doi:10.1039/c3cs60050e
53. International Union of Pure and Applied Chemistry. http://www.iupac.org/. (Accessed Apr 2014)
54. National Institute of Standards and Technology. http://www.nist.gov. (Accessed Apr 2014)
55. Ray JM (ed) (2014) Research data management: practical strategies for information professionals. Charleston insights in library, archival, and information sciences. Purdue University Press, West Lafayette
56. Vaughan KTL, Hayes BE, Lerner RC, McElfresh KR, Pavlech L, Romito D, Reeves LH, Morris EN (2013) Development of the research lifecycle model for library services. J Med Libr Assoc 101(4):310–314. doi:10.3163/1536-5050.101.4.013
57. Coles SJ, Frey JG, Bird CL, Whitby RJ, Day AE (2013) First steps towards semantic descriptions of electronic laboratory notebook records. J Cheminform 5. doi:10.1186/1758-2946-5-52
58. Cuevas-Vicenttín V, Dey S, Köhler S, Riddle S, Ludäscher B (2012) Scientific workflows and provenance: introduction and research opportunities. Datenbank-Spektrum 12(3):193–203. doi:10.1007/s13222-012-0100-z
59. Roadmap for synthesis in the 2st Century Engineering and Physical Sciences Research Council. http://www.dial-a-molecule.org/wp/wp-content/uploads/2012/10/Dial-a-Molecule-Roadmap.pdf. (Accessed Apr 2014)