JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Creating Context for the Experiment Record. User-Defined Metadata: Investigations into Metadata Usage in the LabTrove ELN

Cerys Willoughby,* Colin L. Bird, Simon J. Coles, and Jeremy G. Frey

Chemistry, University of Southampton, Southampton SO17 1BJ, United Kingdom

Supporting Information

ABSTRACT: The drive toward more transparency in research, the growing willingness to make data openly available, and the reuse of data to maximize the return on research investment all increase the importance of being able to find information and make links to the underlying data. The use of metadata in Electronic Laboratory Notebooks (ELNs) to curate experiment data is an essential ingredient for facilitating discovery. The University of Southampton has developed a Web browser-based ELN that enables users to add their own metadata to notebook entries. A survey of these notebooks was completed to assess user behavior and patterns of metadata usage within ELNs, while user perceptions and expectations were gathered through interviews and user-testing activities within the community. The findings indicate that while some groups are comfortable with metadata and are able to design a metadata structure that works effectively, many users are making little attempts to use it, thereby endangering their ability to recover data in the future. A survey of patterns of metadata use in these notebooks, together with feedback from the user community, indicated that while a few groups are comfortable with metadata and are able to design a metadata structure that works effectively, many users adopt a "minimum required" approach to metadata. To investigate whether the patterns of metadata use in LabTrove were unusual, a series of surveys were undertaken to investigate metadata usage in a variety of platforms supporting user-defined metadata. These surveys also provided the opportunity to investigate whether interface designs in these other environments might inform strategies for encouraging metadata creation and more effective use of metadata in LabTrove.

1. INTRODUCTION

Metadata is an essential aspect of the preservation of knowledge for future exploitation, and using metadata to provide context to data that would otherwise be difficult to use is recognized as an important component in the creation of intelligent open data[1,2] for science.

The role of metadata in electronic record curation and electronic laboratory notebooks has been described in detail elsewhere.[3] Discussions about metadata in this context typically focus on the "burden of curation" and other difficulties in metadata generation.[4−9] Capturing experiment metadata in an ELN for "curation at source" mitigates some of the burden of curation.[10] Despite the important role of ELNs in capturing metadata, there appears to be little discussion of how metadata should be used and perhaps even more importantly if and how it is being used.

The Chemical Informatics Group at the University of Southampton has developed LabTrove (www.labtrove.org), a researcher-centric Web- and cloud-based ELN that has a blog-style structure. LabTrove contains a facility enabling users to add their own user-defined metadata to describe their entries and experiments.[3,11] In section 2 of this paper, we report an investigation into how researchers use metadata within the LabTrove ELN and provide insights into what types of metadata are used and whether it is used effectively. Although the LabTrove team had some expectations about how metadata might be used and perhaps ought to be used in the ELN, until the survey we had only a limited knowledge of how metadata was being used by the community.[3,11] We also present some of the feedback and observations about metadata and metadata use that we have obtained from undertaking various user research activities with the LabTrove community. The feedback and observations from these activities raised issues that contributed to the decision to carry out this survey.

In section 3 of this paper, we investigate whether the patterns of metadata use in LabTrove are unusual, using a series of surveys to investigate metadata usage across a variety of other platforms that support user-defined metadata. We end this paper with section 4, with observations from the surveys and a discussion of strategies for improving the use and capture of user-defined metadata for researchers in chemistry based on our findings across both sets of surveys.

2. LABTROVE METADATA SURVEY

LabTrove is a blog-based ELN tool that is used by academic researchers in a heterogeneous set of academic laboratories with groups engaged in analytical chemistry, X-ray studies, drug discovery, and biomaterials research; it is used for a range of purposes such as for recording experiments, data archiving, and

Received: July 22, 2014
Published: November 18, 2014

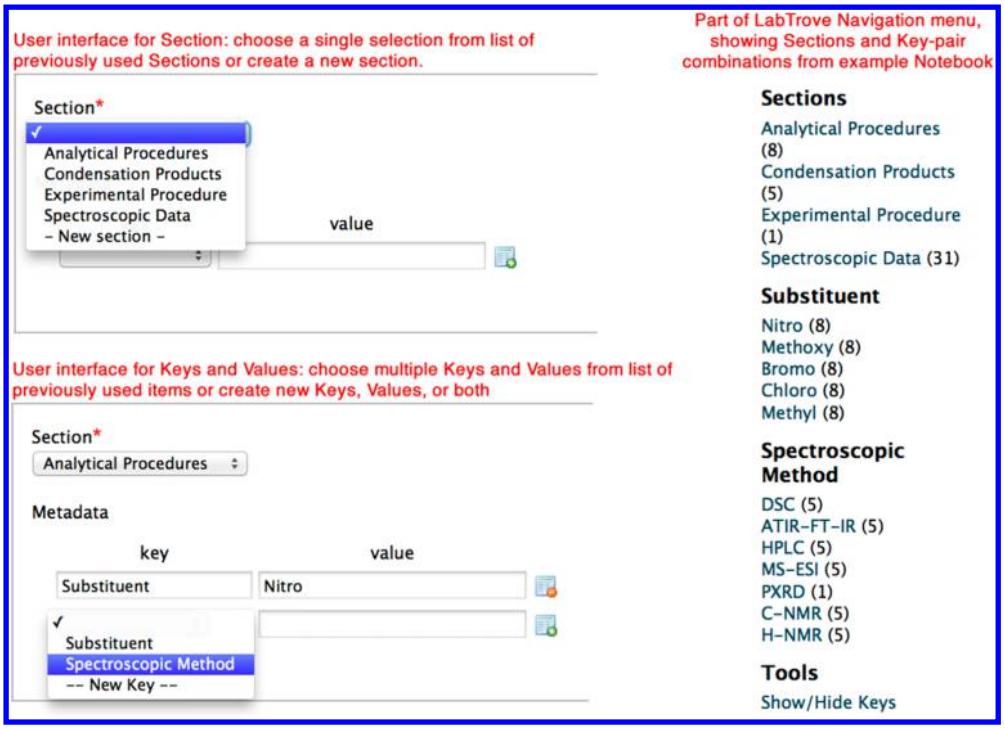

Figure 1. LabTrove interface for adding metadata to a notebook entry and exposure of metadata in the navigation menu (annotated with descriptions of the metadata elements).

project coordination.[3,11,12] Each LabTrove instance, known as a Trove, hosts a collection of notebooks, which may be authored by a single researcher or by a group. Typically, a Trove contains a collection of notebooks belonging to a single research group or a single institution, but some other Troves are shared with notebooks and contributions from researchers from different institutions.[12] Each notebook enables the author to create entries, upload data in the form of files, add user-defined metadata, create links to other materials, and interact with others' notebooks by adding comments.[11] Users can view entries in the notebook in reverse chronological order or can make use of the search facility or a menu generated from added metadata to navigate to specific entries of interest.

Metadata means different things to different people but is usually about more than just "data about data".[9] Metadata can be considered from the perspective of what the metadata describes, such as content, context, structure, preservation, and administration of information,[13] and from the perspective of the function of the metadata, such as allowing resources to be discovered, what the resources are, and how they are organized, which enables use, interoperability, and provides identification and authentication.[14,15] We have previously defined metadata as descriptive information and classification labels that group related items, provide context, and facilitate the reuse of specified research outputs.[16] LabTrove is intended as a "marked up record that can be shared and searched" with the intention that notebook entries record the details of the scientific process together with the data that is produced as a result.[11] The intended functions of the LabTrove metadata features are to describe the context of the research process, enable authors and others to search for and find the research, organize entries by facet, enable reuse both within the ELN and through interoperability with other systems, enable provenance inspection, and archive research. Different styles of metadata and different terms may be more or less effective for each of these metadata purposes.

To investigate the behaviors of research communities and their patterns of metadata usage in LabTrove, a survey of over 100 LabTrove notebooks, covering a variety of disciplines including chemistry, biology, and physics, was conducted.[17] The surveyed notebooks were either entirely public, publicly accessible through an Open ID, or accessible only with an institutional (university) login.

2.1. LabTrove Metadata Elements. The blog-style structure of the LabTrove ELN enables users of the system to record their experiments and activities with individual entries in the notebook. LabTrove provides users with the means to add their own user-defined metadata to their entries in addition to the machine-generated metadata.[11]

The values that are provided for the metadata are utilized internally to create the navigation menu displayed on the right-hand side of the ELN interface. Therefore, the inclusion of metadata makes it easier for both authors and their collaborators to find entries about particular topics or relating to particular experiments. The use of consistent metadata potentially produces a much more effective record than is possible with a paper notebook.[3]

For each entry that a user creates, they are required to specify a value for a Section field. Users can optionally choose to add further metadata to the entries in the form of key-value pairs.[18] The key-value pairs enable the inclusion of metadata that is much richer than could be produced using a simple tagging system and provides a form of classification for the notebook entries. Key-value pairs are used to characterize notebook entries, enable grouping of related experiments, and aid in search and linking.[12] Figure 1 shows the LabTrove user interface for adding sections and key-value pairs and part of the navigation menu.

2.2. Templates. Templates are a special type of notebook entry that can be used for experiments that are repeated or use parallel procedures to reduce the burden of entering the same information repeatedly. These templates are typically used to provide a structured way to enter details of a procedure or experimental results, for example, through the use of tables and structured input fields.[11] Templates can also be configured to encourage users to add consistent and complete metadata by enabling the template author to define metadata in the template. This metadata is then inserted into newly created posts, although users are free to remove or change the inserted metadata and to add their own instead if they want to. To create a template in LabTrove, the author of the entry sets the Section to "Templates". When a user views the template entry, there is an option to "Use Template" that creates a new notebook entry with the structure provided by the template together with any metadata that has been predefined. The user can then modify the contents of the entry, completing any tables or input fields, and making any appropriate changes to the metadata before saving as a normal notebook entry. Despite the advantages of templates to encourage structured and consistent notebook entries, they are not widely employed, and experiences have been mixed. It is relatively easy to create an entry with a standardized structure without using a template, but we are seeing an increasing recognition of the benefits of templates from users and anticipate their use to increase in the future.[12]

2.3. Survey Method. As shown in Figure 1, the metadata elements used within each of the notebooks is present in the navigation menu on the right of the LabTrove interface. To extract the metadata from each of the notebooks for the survey, each notebook was accessed using the LabTrove Web interface and a copy taken of the Section, Key, and Value elements from the navigation menu. Other information was also recorded about each notebook, including the number of authors, number of entries made in the notebook, whether the notebook was primarily used for "autoblogging", and apparent primary use of the notebook. Table 1 shows the total numbers of notebooks, entries, authors, and different metadata elements that were collected from the notebooks and used in this study.

Table 1. Number of Notebooks, Entries, Authors, and Metadata Elements of Each Type Examined in the Survey

notebooks	entries	authors	sections	keys	values
104	22,818	202	512	263	959

Notebooks associated only with the testing or development of Troves were not included in the survey in order to exclude irrelevant metadata. The notebook survey did include several notebooks that were set up with "auto-blogging", where an instrument or sensor automatically enters data to a LabTrove notebook. The notebooks in the survey also included those with templates and entries created from templates but not notebooks only created for the storage of templates. The results therefore include Section metadata with a value of "Templates" and any metadata that was automatically added to entries that were created from the templates. Templates themselves typically only have Section metadata. Although templates can be used across multiple notebooks within a Trove, the majority are actually used within a single notebook. Although some duplication of terms can be expected by template use across multiple notebooks, the results are the same as seen when consistent metadata is consciously adopted by the group[11] as described in section 2.6.3. The metadata examined in the study are investigated by the terms themselves and not by the number of times they are used; so the metadata terms used on entries created by templates, even if a template has been used dozens of times, does not skew the results of the survey.

2.4. Use of Section Metadata. For each entry in the LabTrove notebooks, it is mandatory for users to enter a value for the Section metadata. No default values for Section are provided, but the user can choose to use values that have previously been used in the same notebook or to create a new value. The Section enables a top-level categorization of all entries. Although the LabTrove team did not prescribe any specific uses for the Section metadata,[11] there has been active debate within the team on how Section should be used and what is the best way to divide up entries in the ELN for an experiment. Two alternatives have been proposed:

1. Multiple entries are used to represent an experiment, with different values for Section that match the stage of the experiment or phase of the research process (i.e., Plan, Materials, Procedure, Results, etc.).
2. A single entry is used to represent the entire experiment, and the value of Section is derived from some other aspect of the experiment.

The survey revealed that a very limited range of values are being used for the Section metadata, with 70% of the notebooks using five Sections or less and around half of those using only one Section; only 10% of notebooks use more than 10 different values for Section. The most frequent type of word or short phrase used as metadata for Section was a "catch-all" phrase, for example, the Section has a blank value such as a space or a dash or only one Section is used with an imprecise value such as *General*. Almost half of the notebooks use a catch-all Section of some sort, with 35% of notebooks using a catch-all Section in addition to other Section values. This frequent use of a catch-all Section, coupled with high numbers of users using only a single Section suggests that that a large percentage of the notebook authors are avoiding adding metadata and are taking a "minimum required" approach, where metadata has only been added because the system requires it.

2.5. Use of Key-Value Metadata. The author of an entry in the ELN can optionally add metadata in the form of key-value pairs. For example, the user may choose metadata to add an experiment ID to their ELN entries. In this case, the chosen Key could be "Experiment ID", and the value could then contain the identifier for that experiment. There is no limit to the number of key-value pairs that can be added to an entry, and each Key can have multiple values. Key-value pairs are not just limited to entering specific identifiers in this way but can be used both to capture the characteristics of individual entries at a variety of levels and to provide the capability to categorize entries using subcategories that are relevant to the individual author. Where the Section metadata key enables a top-level categorization of all entries,[11] the key-value pairs provide a form of two-level classification that complements the flexibility of the notebook entries.

The number of metadata Keys used was counted within each notebook. The figures show that almost half of all notebooks used no Keys at all, with only about a quarter using three or more Keys, indicating that this powerful metadata capability is poorly used within the community. The lack of examples of

good practice for this type of metadata probably contributes to this lack of use.

2.6. Classification of Metadata in LabTrove. To investigate the metadata in more detail, the metadata terms were classified in a number of ways. Each metadata item was classified by determining whether the term could be considered high-level or specific, that is, whether the metadata used in the ELN is typically generic or more specialized. More specialized metadata is likely to be more useful for locating entries about specific topics, whereas high-level terms are likely to be applicable to more entries and potentially less useful. The words "Data" or "Enzymes" would be classified as high-level terms, while "Filament" or "Herbal Medicine" would be classified as specific metadata. Figure 2 shows more examples of Sections and their categorization into High-Level and Specific groups.

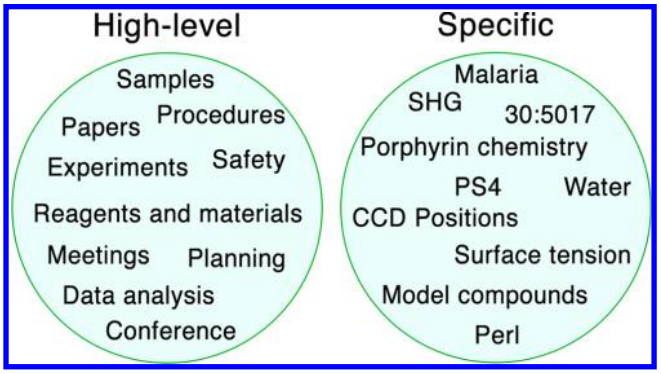

Figure 2. Examples of Section metadata classified as "High-Level" or "Specific".

The LabTrove metadata was also classified by word or phrase type into Noun-type, Verb-type, and Adjective-type. For single words, this is straightforward with the words classified by their dictionary class. For ambiguous words and phrases, the most appropriate class was selected. For example, the terms "Platform management meetings", "Electrochemical properties", and "Interesting new papers" are classified as noun-type phrases. "Freshly crushed", "shade of blue", and "math-related" are classified as adjective-type phrases. "Data (formatting)", "Learning design", and "Electrochemical cleaning" are classified as a verb-type phrases.

Finally, the Noun-type metadata items were classified based on the dominant subjects for each metadata type, as determined by an initial coding exercise. The resulting categories are similar, for each metadata type, with Activities, Codes, Dates and Values, Equipment and Instruments, Labels, and Materials present for each type. Some differences are observed. In particular, a Catch-all category is only seen in the Section metadata, and a Location category is only observed in the Value metadata.

2.6.1. Metadata Is High-Level Rather than Specific. The results of the survey showed that the majority of the metadata used could be classified as high-level rather than specific, with more than two-thirds of the Section metadata and 80% of the Key metadata classified as High-Level. This finding suggests that the capture of metadata may not have been very effective because specific metadata that would help to provide context and identification for individual experiments or records is missing.

2.6.2. "Things" Are More Significant than "Activities". The results of the classification exercise indicate that high-level labels, or descriptions of objects or properties such as materials, data, or instruments, dominate the metadata terms used. The vast majority of metadata terms are Noun-type, including dates, sample numbers, and instrument locations, with Adjective-type often used for the catch-all metadata terms such as miscellaneous and general. Less than 5% of the metadata items are Verb-type, such as modeling, blogging, or monitoring. Nouns that describe activities, such as "purification", "preparation", "analysis", "lithography", "electrophoresis", and "filtration", are used at least twice as often as verbs to represent activities, but even so, these types of nouns that describe techniques, methods, and actions used to complete the experiments are still relatively poorly represented in at less than 10% of the total metadata. There are a couple of potential hypotheses that might explain this observation. It may be easier (or more natural) to classify notebook entries by physical properties or objects rather than activities or it could be that the activity is considered to be less important to record, but the reasons were not investigated.

Table 2 provides a detailed breakdown of the word-type classifications for the different metadata types. The word-type classifications are used to compare the LabTrove metadata with other platforms later in this paper.

Table 2. Breakdown of Word-Type Classifications for Metadata

	Noun-type	Verb-type	Adjective-type
Sections	513	15	3
Keys	263	4	1
Values	893	40	26

2.6.3. Section and Key Metadata Are Dominated by High-Level Labels. The results of the subject classification exercise show that the majority of the Section metadata are high-level labels representing the content of the entries that they characterize, while the remainder describes items such as codes, dates, instruments, materials, topics, and project information. Figure 3 shows a word cloud of Sections classified as "Labels". Similar terms have been combined, for example, material and materials. As mentioned previously, templates are a specific type of notebook entry in LabTrove for which users are required to use the value "templates" for the Section. Of the

Figure 3. Wordle (http://www.wordle.net) visualization showing Section metadata classified as "Labels". Size relates to the relative frequency.

dx.doi.org/10.1021/ci500469f | J. Chem. Inf. Model. 2014, 54, 3268−3283

surveyed notebooks, 28% contain templates, and as such, it represents the most frequent Section term used.

The majority of Key metadata are also high-level labels with other Key terms representing items such as specific instruments, materials, documents, and properties. Figure 4 shows a

Figure 4. Wordle visualization showing Key metadata classified as "Labels". Size relates to the relative frequency.

word cloud of Keys classified as Labels. Similar terms have been combined, for example, data type and data. It can be seen that very similar words and phrases are used for the Section and Key metadata, suggesting that these terms are particularly important for recording experiments. This duplication across the Section and Key metadata also shows that users are not differentiating between the metadata types effectively.

Figure 5 shows the categorization of the Values metadata in the LabTrove entries and shows that these metadata refer

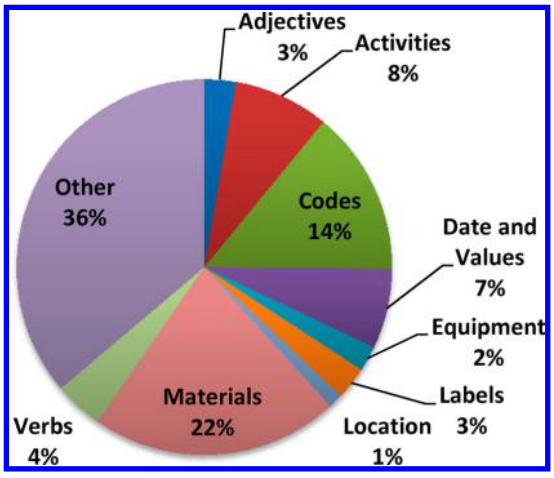

Figure 5. Categorization of Values metadata.

primarily to materials, activities, and various codes, such as notebook and experiment identifiers, dates, and activities. Examples of Values metadata categorized as "Other" include "Alpha", "computational science", "E-notebooks", "Forms", "Introduction", and "Matlab".

The "post-type" Key name is used primarily, but not exclusively, by a set of notebooks related to bioscience experiments. The group conducting these experiments has consciously made a decision to use "post-type" as a defined a set of values associated with the Key, including material, data, sample, procedure, safety, template, and note.[11] This group is in fact the main group that uses templates across multiple notebooks, but the "post-type" Key and its associated value, describing the type of information captured by the entry, is typically the only automatically added metadata, while entries not created from a template are assigned an appropriate "post-type" value or Section by the author. Additional Sections, "post-type" values, and other key-value pairs are assigned personal terms by individual authors based on their own needs.[11]

A full list of the metadata categories and numeric breakdowns are provided in the Supporting Information.

2.7. Impact on Metadata of Privacy and Number of Notebook Authors. The majority of notebooks in the survey (65%) have only a single author, but 8% of notebooks have five or more, with one notebook having 25 authors. The number of authors on a particular notebook appears to vary by the function of the notebook, with a higher average number of authors seen in notebooks where the primary function is recording group and project activities (2.95) or for discussions about the ELN itself (2.46) compared to notebooks primarily used for recording experiments (1.40). The majority of notebooks with two or more authors are notebooks available only through institutional logins. The authors on these notebooks are typically collaborators belonging to the same group and institution, but even for the publicly available notebooks, the majority of authors are co-located and belong to the same group.

The highest numbers of Sections are observed in notebooks with the highest numbers of authors. There is also a trend toward increased key-value metadata use with an increase in notebook authors. There appear to be two reasons for this increase depending upon the function of the notebook. In notebooks with multiple authors, which have relatively high numbers of Sections or Keys, and that are used primarily for formal recording of experiments or project activities, the metadata terms used relate to many different topics, with no duplication of meaning. For those notebooks that are used primarily for discussing and experimenting with the ELN itself, or "sandpit" notebooks, the metadata terms used are much less broad and often include terms that have the same meaning or are very similar, for example, "procedure" and "procedures", "API testing" and "testing API", "test" and "testing". Discussion with users has also highlighted a problem where multiple authors use the same metadata term but may use different capitalisation leading to different metadata stored in the notebook. The example given was the term "NMR" used as a Key, where different authors had entered "NMR", "Nmr" or "nmr", leading to multiple Keys being created. Although removing case sensitivity is one option, a merging process to replace or standardize metadata values may be useful where duplication occurs, such as with abbreviations, synonyms, homonyms, and misspellings.[19−21]

The notebooks with the highest numbers of authors also have a higher number of entries, which may in turn lead to the creation of more metadata to describe the additional activities. When the number of entries in a notebook is high, it may be more apparent that better organization is required, which would drive the production of metadata in those notebooks. This is supported by the broader range of terms used for metadata in these notebooks when they are used for recording experiments and organizing project work.

Interestingly the notebooks with more than five authors appear to rely more on Sections than on key-value pairs. There could be a number of reasons for this. For example, as the majority of these notebooks are authored by members of the same group, there may be less of a need for detailed metadata, although several groups mention that they make use of key-value pairs in order to help enable grouping of entries.[12] An

alternative reason may be that large numbers of key-values are quite difficult to manage in the current user interface and may actually hinder the rapid location of entries.

Are notebook users adding metadata for their own benefit or for that of readers? The owner of a LabTrove notebook can choose the level of privacy for their notebook. The privacy level of the notebook may indicate if the author intended it to be used for personal record keeping or data dissemination and therefore whether the needs of an audience may have been taken into account.[22] If the author of a notebook intends their work to be read by the public or peers, then including more metadata will enable their work to be found and used more easily. Authors ought therefore to choose to add more metadata in the ELN, although that is no guarantee that it is created to meet the needs of an external audience.[23]

The notebooks included in the survey were divided into three groups based upon their privacy settings:

1. Entirely public (Public)
2. Publicly accessible through an Open ID (Logon)
3. Accessible only with an institutional (university) login (Private)

Notebooks that are only visible to the notebook owner or specifically selected users do not form a part of the survey. The numbers of metadata elements used were compared for each group to determine whether privacy did have an effect on the amount of metadata used. Figure 6 shows that the results are

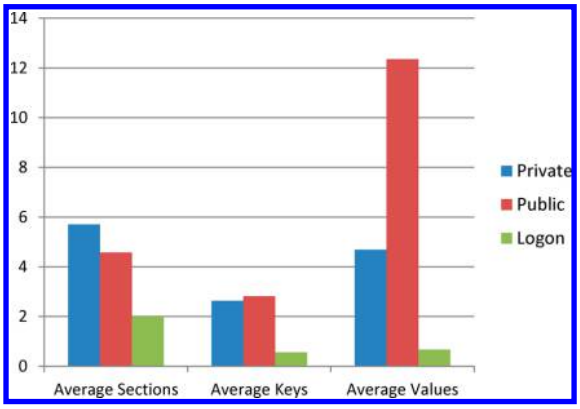

Figure 6. Average number of Section, Key, and Value metadata by the privacy status of the notebook.

somewhat unexpected, with the Logon group of notebooks showing relatively little metadata at all. Looking into this in more depth, the notebooks in the Logon group are not well populated, and those with low metadata numbers are notebooks for discussion by a specific community rather than for formal experiment recording or engagement with the general community. Another notebook is for experiment recording, but it is populated through autoblogging. So the metadata is the same for each entry. These notebooks need to be accessible to users from a variety of different institutions, and this is likely to be the reason why the particular privacy settings have been chosen.

Given that that one of the perceived inhibitors to metadata use is the reluctance to make data public in the first place, it is perhaps surprising to see that the private notebooks have some of the highest average figures for metadata use. However, many of the notebooks that are protected behind institutional logins are intended to be public within their own community.

One of the groups with the largest number of publicly accessible notebooks has expressed anxieties about creating metadata in their notebooks but have recently begun to add more metadata, particularly key-value pairs, to help them organize and group their experiments.[12]

A further study would be needed to determine whether the metadata values contribute to community engagement or help with discovery of experiments and notebooks. LabTrove has a search facility that includes the metadata related to entries, as well as the entry titles, and entry contents. A possible extension to LabTrove in the future could be to provide statistics and data analytics in order to investigate what terms are used for search, how users find a particular notebook or entry, and which entries are popular. From discussions with users, we know more about what the authors of notebooks search for within their own notebooks than we do about how other users use the search facility or what information those users are trying to find. Interviews with users indicate that users primarily use the Navigation menu to locate notebook entries of interest, using all of the different metadata elements that are available including dates, authors, Sections, and key-value pairs. Users also use the search facility, but this is typically to find specific experiment identifiers and can often be many months after the original experiment was recorded. For some of the groups a new researcher may continue work on a set of experiments and therefore take over recording in a notebook for a researcher that has moved on. Researchers in this position have indicated that better metadata would have helped make their job of taking over research easier.

For the moment, the number of authors remains a better indicator of community engagement until statistics on search and the number of unique visitors to a notebook become available.

2.8. LabTrove User Research. The LabTrove development team have undertaken a variety of activities to investigate user behavior and attitudes toward metadata including interviewing both current users and communities interested in adopting LabTrove, usability testing with novice users, and trialling the ELN with students. These activities have provided the opportunity to examine what expectations and understanding users have about metadata and how the design of the ELN might affect their metadata use.

The results of these activities have indicated that although some users are comfortable adding metadata to their notebooks and understand the benefit the metadata provides in helping them to locate information, many others have indicated that they felt it was too difficult to use and did not see the benefit in adding it. A great deal of the anxiety within the user communities about using metadata stems from a fear of having to design their own metadata scheme. The researchers want meaningful examples to follow but feel that taking schemes from other teams would not be very helpful because the scheme would be too specific and not easy to adapt. The alternative of being provided with an example created at a high level would be too generic to be useful. Even though the researchers do not know exactly what metadata they want to use, their reluctance to use schemes from other teams may indicate that they do nevertheless have the ability to recognize a meaningful example when they see one.

Observations from our user studies and the ELN trial with students match the results seen in the metadata survey in that the majority of new users add the minimum amount of metadata required (a single Section). In common with the data

from the notebook survey, if adding metadata were not enforced within the notebook, the majority of users would not add it, at least with the current interface design. Students differed in how they approached writing entries in the ELN, but the way they chose to structure their experiments affected how they used metadata. If they chose to break the parts of their experiment into multiple entries, they were more likely to use appropriate metadata to classify their entries by content or activity, whereas those students who created a single entry for an experiment tended to use a catch-all category. Where example Section and Key terms were provided, for example, from exemplar notebooks or template entries, the students were more likely to use metadata, but they did not necessarily select appropriate combinations or contribute their own more relevant terms.

Several of the interviewed users indicated that they were more familiar with tags from using other Web-based tools and felt more comfortable adding metadata as tags. In fact some of the researchers described the metadata that they use as "metadata tags" and wanted to have the metadata visually represented in tag cloud form. There is evidence that social networking is one of the main areas where students encounter and actively use metadata.[24] Some of the users also expressed a desire for different ways of organizing their data in the ELN, a feature that could be facilitated if users were encouraged to provide specific metadata such as the start date of the experiment or sample identifiers. Others viewed metadata as a way of making links to other resources, suggesting that users expect metadata interfaces to behave in a similar manner to familiar social networking applications such as Facebook or Flickr, where information becomes a link to other people or objects that share the tag or name.

Some anxiety about using metadata comes from a fear of using the "wrong" metadata and therefore having to go back and change it, or "messing up" data by creating mismatched metadata between entries. Some of the groups that were reluctant to use metadata in their notebooks for these reasons have more recently begun to add metadata to their notebooks entries and to view metadata as a useful tool for grouping experiments together.[12] Adding metadata retrospectively is not difficult once a suitable scheme has been devised.

3. INVESTIGATING USER-DEFINED METADATA USE ON OTHER PLATFORMS

Although the LabTrove team had some expectations about how metadata might be used and perhaps ought to be used in the ELN, until the survey, we had only a limited knowledge of how metadata was actually being used by the community.[3,11]

The results of the notebook survey provided insights into two main areas. First, that of the utilization of the metadata facilities provided, where it was found that although some groups are comfortable with metadata and are able to design a metadata structure that works effectively, a large percentage of the community adopted a "minimum required" approach, where metadata has only been added because the system requires it. It was also found that metadata use increased with the number of authors for a notebook, but the amount of metadata used was not associated with the privacy status of the notebook. This is not entirely in contrast with studies of other platforms where users can define their own metadata, which suggest that users are often motivated to add user-defined metadata both for social reasons and for personal organization or future retrieval.[25,26]

Second, the survey provided insights into the type of metadata that users were choosing to add to their notebook entries. The results indicate that the majority of the metadata used was high-level rather than specific and is dominated by high-level labels representing the content of the entries that they annotate. The specific metadata describes objects or items such as codes, dates, instruments, materials, topics, and project information. Activities such as techniques, methods, and actions that were used to complete the experiments are relatively poorly represented in the metadata. The vast majority of metadata values are Nouns or Noun-type values, including those that describe activities such as purification, preparation, analysis, lithography, electrophoresis, and filtration, with adjectives and verbs rarely used.

The insights from both areas, utilization and type of metadata, together with feedback from our user community, provide suggestions for changes that might be made to the LabTrove interface to encourage more metadata creation and to improve the quality and usefulness of the metadata that is captured. The results of the survey also raise an important question: Is the metadata used in LabTrove significantly different from other platforms that enable the creation of user-defined metadata? For example, do those other platforms suffer from the same problems of minimum required use, does the number of authors influence the amount of metadata, and are the same kinds of metadata values being used?

Our user research indicated that many researchers have experience with social media and networking platforms such as Facebook and Flickr (section 3.1) that make use of metadata, and presenting familiar interfaces like these might support and encourage the creation of metadata in an ELN. Our examination of other platforms that enable user-defined metadata creation also provided an opportunity to investigate whether particular interface designs might be effective at encouraging user-defined metadata creation. These platforms include Flickr (section 3.1), blogging platforms used by NASA (section 3.2) and other researchers (section 3.3), and myExperiment (section 3.4).

In this section of the paper, we present the results of surveys of user-defined metadata use across different platforms to investigate if the metadata usage seen in LabTrove is unusual and whether the types of metadata used are different across different communities and platforms. We use the results of these investigations to discuss potential enhancements that could be made to improve metadata creation, capture, and quality in LabTrove (section 4).

3.1. Flickr. Flickr is an image sharing social networking site, where users can add tags to describe their photos and videos. The tags are displayed reasonably prominently on the pages displaying the images, together with an invitation to add tags. Interfaces for uploading images on the Web site and also through mobile applications usually have a prominent invitation to add tags, and the user can select from their commonly used and previously used tags. Other Flickr users can also tag the images. This tagging by other members has been used effectively by organizations such as the National Library of Congress and the British Library to increase access to publicly held photography collections and to acquire information and knowledge about the subjects in the photographs from the general public through "The Commons" project.[27] There are also "machine tags"[28] in Flickr that use a special three-part syntax. Machine tags are frequently added by third party Flickr applications to provide information about themselves and

geotag photos and for more novel uses such as the Proyecto Agua that uses the machine tags to add metadata about the taxonomic classification of waterborne organisms.[29]

Although it is difficult to gauge what percentage of users use tags in Flickr, of one of the author's (CW) 657 contacts, less than 4% used no tags at all, with an average of 12 tags on each picture. Less than 10% use 20 or more tags.

For the Flickr survey, Flickr Hive Mind,[30] a data-mining tool for the Flickr photography database, was used to acquire tags from the 500 images ranked most interesting that were tagged with both "chemistry" and "experiment".[31] This produced 1381 unique tags with a further 162 removed because they were indecipherable. These tags were then classified by word or phrase type and by topic in the same manner as the LabTrove metadata. Figure 7 shows the breakdown of the Flickr tags by

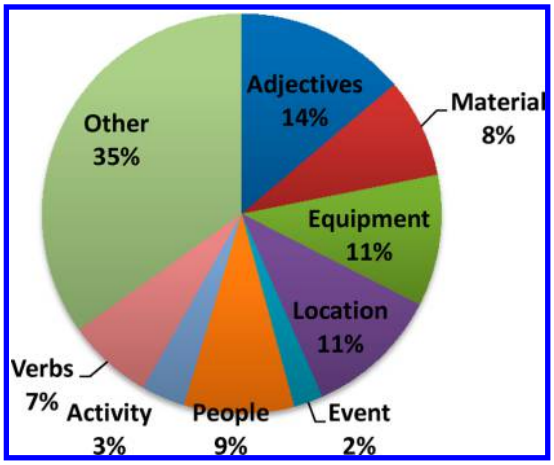

Figure 7. Breakdown of tags on Flickr images tagged with "chemistry" and "experiment" by word type and subject.

word type and subject. It can be seen that Verb-type usage is slightly higher in Flickr (7%) compared with LabTrove (3%), while usage of the descriptive Adjective-type is much higher in Flickr (14%) compared to LabTrove (2%).

According to the tags in Flickr, materials and equipment are clearly important in chemistry experiments, with location and people also common topics of tags. The locations are mostly specific locations, such as specific institutions or towns, but with some more general locations such as "library", "school", and "laboratory". These values are similar to location values used in LabTrove, which include a mixture of specific towns and specific laboratory locations. Metadata about people are more common in Flickr than LabTrove, with mostly general values such as "person", "man", "physicist", and "researcher", but also some specific values such as the names of individuals.

3.2. NASA Blogs. NASA blogs might not seem to have very much in common with chemistry notebooks, but the system that is used by a number of NASA organizations and personnel has some parallels with the LabTrove system. The NASA blogs are based on the WordPress platform.[32] The majority of the NASA blogs use tagging, and it appears that entries are required to have at least one value.

Of the 53 NASA blogs investigated, only 30% of the blogs use three or less metadata items, although nearly 80% of the blogs use the "General" tag. It is unlikely that users are choosing to add the "General" tag deliberately due to the location of the tagging feature far down the interface in the NASA blog system, and therefore, the "General" tag appears to be automatically added to entries when no other tags are included.[32] The inclusion of a default metadata value is normal in WordPress blogs, although it most commonly has a value of "uncategorized". Around 50% of the blogs have multiple authors, including one blog that has 45 authors!

User-defined metadata from the 53 NASA blogs was extracted from tag clouds and category lists to generate 1226 items of metadata. These metadata items were then classified using word or phrase type and subject, as shown in Figure 8.

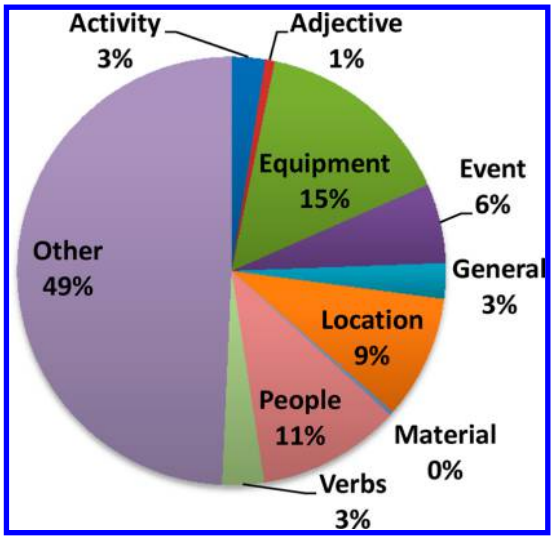

Figure 8. Breakdown of metadata on 53 NASA Blogs by word type and subject.

Full details of the NASA blogs, including names, URLs, extracted metadata, and classifications, can be found in the Supporting Information. The results show that NASA blogs use verb- and adjective-type metadata even less than LabTrove, with the metadata dominated by equipment, such as various rockets and instruments; locations, particularly NASA sites; and people, such as astronauts and NASA leaders, organizations that NASA works with, and the general kinds of people that NASA programs are targeted at, such as women and students.

Figure 9 provides a word cloud of the metadata from 53 NASA blogs, showing clearly how the metadata represents the

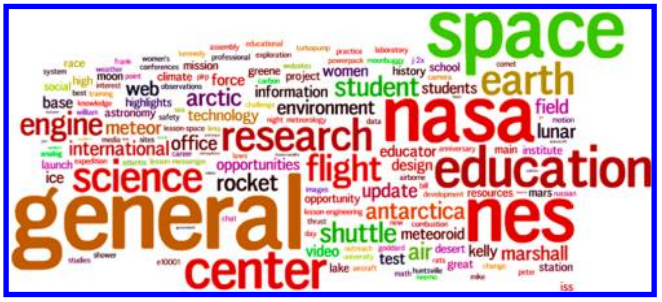

Figure 9. Wordle visualization showing metadata from 53 NASA blogs. Size relates to the relative frequency.

interests of the NASA organization, and also the frequency of the "General" tag.

3.3. Chemistry Blogs. The comparison to the NASA blogs is interesting from the point of view of user behavior and word type usage but less relevant in terms of subject classifications of the metadata. In order to address this limitation, metadata from

50 chemistry-related blogs on the WordPress and Blogger platforms were investigated. Both the WordPress and Blogger platforms encourage users to add metadata to their blog entries in the form of tags and categories on WordPress and "labels" on Blogger. Whether the metadata is visible on entries or within the interface in tag clouds or other navigation assistants varies depending on the blog author's settings. WordPress has an interesting feature that encourages the user to add tags to a newly created entry based on the tags that others have used, but this seems to be using only other tags that have been added rather than making suggestions based upon the content of the entry.

The chemistry-related blogs in the study were found through searching Technorati or from recommended chemistry blog compilations. Only blogs that had accessible metadata were chosen for the investigation, with a further 90 blogs excluded from the study because they had little or inaccessible metadata. One of the advantages with the Blogger platform when extracting metadata is that you can get a list of the available labels by adding a query string to the Web address of the blog, whereas extracting metadata from a WordPress blog relies upon a tag cloud or category list being visible in the chosen theme.

User-defined metadata from the 50 chemistry-related blogs were extracted from tag clouds, category lists, and label lists to generate 10,436 items of metadata. These metadata items were then classified using word or phrase type and subject, as shown in Figure 10. Full details of the chemistry-related blogs, including names, URLs, extracted metadata, and classifications, can be found in the Supporting Information. The results of the classification are very similar to the NASA blogs in that the number of adjective- and verb-type words or phrases is relatively low, and "things" represent the majority of the metadata, as seen in the other metadata sources. When the metadata is classified by subject, the dominant category comprises values that describe materials such as organisms, chemicals, and drugs. Categories describing equipment, people, and locations have smaller numbers of values. One third of the blogs use a "catch-all" category such as "uncategorized", the default if no metadata is chosen in WordPress, or values like "General" and "Miscellaneous". Figure 11 provides a word cloud of the most common metadata values from the chemistry-related blogs, showing the variety of different topics covered by the blogs. The word cloud also demonstrates how the high-level type metadata becomes more visible and therefore appears more important than the specific terms in this kind of visualization.

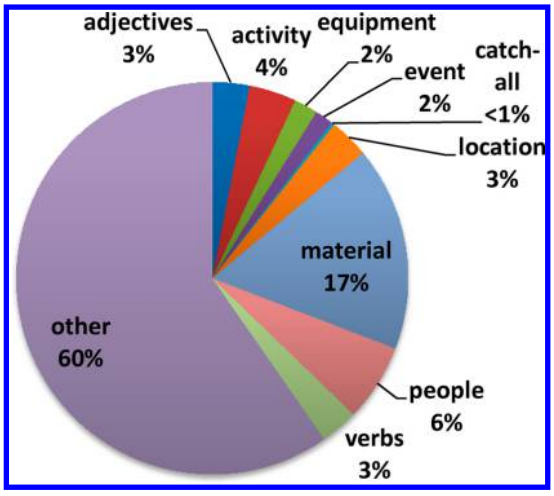

Figure 10. Breakdown of metadata from 50 chemistry-related blogs by word or phrase type and subject.

Figure 11. Wordle visualization showing metadata items from 50 chemistry-related blogs. Size relates to the relative frequency.

3.4. myExperiment. The myExperiment Web site[33] was chosen for surveying metadata because the primary function of the Web site is for researchers to store and share particulars relating to their research methods and experiment workflows.[34] We assumed that these particulars are likely to contain a lot of information about research activities and may therefore be tagged with user-defined metadata that reflects these activities, providing a useful comparison to the experiment data in LabTrove.

The myExperiment Web site enables users to add "tags" to describe their workflows. Users can choose from their previously added tags, which are visible in the interface. An interesting addition to the myExperiment interface is the "This Workflow has no tags!" message if the user has not added any tags to their workflow. The message may be effective at encouraging users to add tags by highlighting the workflows where they have been omitted. A sample of 10% of the most recently added myExperiment workflows suggests this tactic would be successful, with 96% of the workflows including at least one tag.

User-defined metadata from myExperiment was extracted from the "All tags on myExperiment" page[35] to obtain 2349 unique items of metadata. These metadata items were then classified using word or phrase type and subject, as shown in Figure 12. The extracted metadata and classifications can be found in the Supporting Information. The results show that

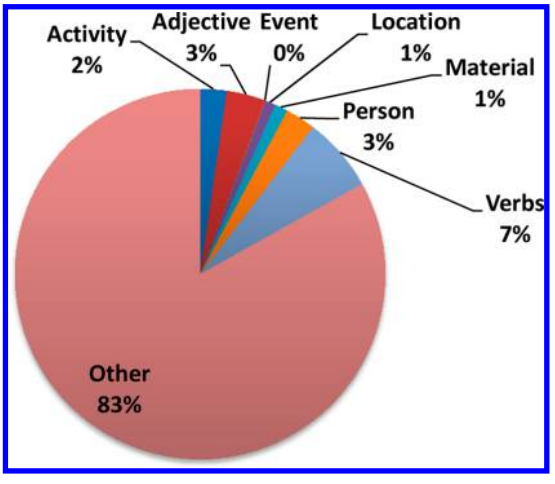

Figure 12. Breakdown of myExperiment tags by word or phrase type and subject.

dx.doi.org/10.1021/ci500469f | J. Chem. Inf. Model. 2014, 54, 3268−3283

despite the site being primarily about activities and processes such as experiment workflows and methods, the tags are still dominated by noun-type words and phrases, with similar percentages of verb-type and activity metadata as LabTrove. The metadata in myExperiment is in fact dominated by computer-related subjects, with over one-quarter of the tags comprising abbreviations and acronyms. Materials, equipment, location, and people are relatively insignificant for this community, which is dominated by computer scientists, information professionals, and researchers.

4. ENHANCING METADATA CAPTURE IN LABTROVE

This section provides some general observations and suggestions for how the user interface and other aspects of the metadata provision might be enhanced in LabTrove based on the findings from the initial LabTrove metadata survey, user feedback, and investigations presented in this paper. The range of suggested enhancements to LabTrove consists of the following areas:

- Development of standard schemas
- Use of invitations to add and reuse metadata
- Automation and automatic assistance
- Experiment and other notebook entry profiles
- Visual benefits to adding the metadata
- Educating researchers

It should be noted that there is no implied commitment by the authors to implement these recommendations in future releases of LabTrove. Many of the observations and suggestions in this section will be applicable to other ELN platforms or interfaces designed for recording experiment and research information.

4.1. Observations from Studies. In this section, the high-level observations from the studies are presented together with how they might influence the design of LabTrove and, in particular, the user interface.

4.1.1. Mandatory Metadata, Default Values, and "Blank Canvas" Effect. The clearest result from the LabTrove notebook survey and feedback from user experiences was that the biggest inhibitor to adding useful metadata is the "blank canvas" effect, where the users may be willing to add metadata but do not know where to start. In some cases, users do not even know what metadata is. Providing example metadata in the form of standard schemas would help users overcome some of their difficulty with getting started in a notebook, but the survey results indicate that making the addition of metadata mandatory without such assistance is not helpful at encouraging the creation of meaningful metadata. In LabTrove, a Section is mandatory, but a large percentage of users use a meaningless "catch-all" value for their entries. Default values are added automatically in the NASA logs and the WordPress-authored Chemistry blogs if no user-defined value has been added to the entries. The majority of the NASA blogs surveyed included the default value "General", and many of the Chemistry blogs included the WordPress default value "uncategorized". The presence of a default value may encourage some users to replace it with something more meaningful, but its prevalence in the survey results suggests that this is not very effective. Overcoming the lack of knowledge about what metadata is and how to use it requires appropriate education about data management and metadata facilities.

4.1.2. Different Communities Use Different Terminology. The results presented in this paper show metadata is representative of the community that uses it, and therefore, no one set of terminology is likely to work for all users of an ELN. For example, "Materials" is one of the most commonly used metadata values in the notebook survey, but the terminology (and types) used for the materials included in an experiment does differ significantly depending upon discipline and the type of experiment. For example, "Inputs", "Reagents", "Strain", "Products", "Substance", "Compound", "Molecule type", "Batch", and "Sample" might be appropriate Keys depending on whether the community is involved in biology, chemistry, or drug discovery.[36,37] However, the results do show that commonalities exist at a high level between these different communities, particularly the "label"-type information such as materials, locations, data types, people, events, and equipment, but also specific topics such as "science", "chemistry", "education", "pharmaceuticals", "technology", "environment", and "research". Selecting the correct terminology in the interface is important for ensuring the relevance of the metadata for the user and to encourage subsequent capture. Basic schemas could be developed using these common elements and included in the ELN, while different communities can develop their own extended schemas, which can be shared and imported into the ELN for seeding by example. As mentioned in section 2, templates are a mechanism that can be used to encourage users to enter experimental information with a consistent structure and also using consistent vocabulary for the metadata.

The results of the surveys also indicate that users more readily add certain types of information as metadata. For example, the majority of metadata used describes things and objects, rather than activities, indicating that some information that would be useful is not captured as metadata, such as experiment activities and conditions. Mechanisms could be provided to specifically seek addition of this useful but underrepresented context for the experiment record.

4.1.3. Invitations. The majority of the platforms examined in the surveys included interface designs with invitations to add metadata. For example, in Flickr, the user is invited to "Add a tag" and "Choose from your tags". In WordPress, the user is invited to "Choose from your most used tags". In myExperiment, the user is invited to "Add tags". Other platforms that users may be familiar with also use invitations to encourage the creation of information. For example, in Facebook,[38] the user is encouraged to add people, locations, emotions, activities, feelings, and items into their status updates and photos with questions such as "What's on your mind?", "Where was this taken?", and "Who are you with?". Another example, LinkedIn,[39] also uses questions inviting users to add information about their career history and interests. None of the platforms included in the study have invitations that are as prominent as those seen in Facebook or LinkedIn, but many of the mobile applications for Flickr include more prominent invitations to add tags than the Web interface. Making the invitation to add metadata more inviting and more prominent in the LabTrove interface is likely to increase the amount of metadata that is added to entries.

4.1.4. Making Use of Metadata. In Flickr, any member's tags can be viewed as a list, and tags can also be used in advanced search. In myExperiment, all tags can be viewed as a tag cloud, and selecting multiple tags in a filter can be used to narrow a search for workflows. In the WordPress and Blogger blogs, the user can choose through the use of themes or widgets how their blog is displayed and can choose to enable the

metadata to be used as a navigation menu or displayed as a list or tag cloud. It is possible for a user to have their blog configured in such a way on these platforms that the metadata is not visible to an audience. This study is unable to comment on the effectiveness of these approaches, but if we want to encourage users to add metadata, then metadata ought to be prominent and provide useful functionality, such as tag clouds, filtering, and the searching capabilities seen across the platforms. In LabTrove, the metadata added to notebook entries is used to create a navigation menu and is also used to provide advanced search functionality, but given that the capability to add user-defined metadata is a strength of the platform, then it should be possible to make better use of the metadata. As discussed in the following sections, there are strategies we can implement to encourage users to add more metadata and to help users to create "better" metadata. Users can be educated to understand how metadata can be used to help them in their day-to-day work, such as finding previous experiments, characterizing notebook entries, grouping entries, and enabling link creation,[12] as well as why it is important for long-term preservation. Better understanding of the benefits of the metadata together with systems and tools making use of this metadata are beneficial for all users.

4.2. Encouraging Addition of Metadata. Capturing metadata for experiments within the ELN provides the opportunity to make metadata creation and curation easier and simpler.[10] As demonstrated by the LabTrove survey, just providing a mechanism to add metadata is not sufficient. More assistance needs to be provided to users to help them create more metadata but also more consistent and appropriate metadata for their experiments. Schemas and ontologies can be used for providing a valuable starting point and consistent terminology and for interoperability with other systems. Users can be supported through more effective interface designs and automatic assistance for metadata creation, and making use of the metadata through visualizations and enhancing the user interface can help users to get value from adding metadata in their every day work.

4.2.1. Providing a Starting Point with Data Using Schemas. The most effective way of solving the "blank canvas" problem is likely to be developing basic schemas, starting with a single generic schema that can be extended by discipline and then extended further to meet the needs of individual research groups. These schemas could be used in LabTrove to provide users with meaningful options for Sections and key-value pairs that they can select when creating their experiments.

Functionality could be provided in LabTrove to enable users and administrators to create and include basic or more complex schemas into LabTrove. The schema could optionally be available in new notebooks. In the United Kingdom, efforts to harmonize subject classification across the higher education and research sector have resulted in the production of a three-layer classification scheme that could provide a basis for metadata schemas for the ELN.[40] The classification includes different discipline areas, methodologies, instruments, material types, and topics.[41] Where metadata is well used, notebooks include terms that match some of the harmonized subject classifications. Individual communities could develop extended schemas and import these schemas into the LabTrove interface. Schemas could be shared with other groups and adapted as required.

Although there is a plethora of standard schemas and ontologies for science, both generic and domain specific, that could be integrated into LabTrove, there is no single adopted format or model. A benefit of integrating such standards into LabTrove is that they provide the opportunity for the exchange of data as well as merely promoting consistent use of metadata terminology. For example, some recent projects by members of the LabTrove communities have created structured data for sharing between the ELN and different systems based on standard metadata requirements.[12,42]

The problem with the myriad of alternatives is choosing which ones to use. Most of the schemas that have been defined also focus on the structure of data rather than the experimental process as a whole, although notable exceptions exist, for example, ORECHEM,[43,44] CMCS,[9] SEML,[45] CCLRC Scientific Metadata Model,[46] SciPort,[47] and CMO.[48] Other questions to be answered include the following: To what extent the schema should be specialized? How to agree common terminology across disciplines? How much flexibility is permitted within the system to add extensions or customize the metadata to your own use? These questions are beyond the scope of this paper. Milsted et al. have provided more in depth discussion of ontologies and LabTrove.[11]

Formal documentation of schema elements would be valuable to ensure understanding and enable consistency of use. Example notebook entries with metadata included should also be provided with the documentation.

In LabTrove, any schema can take advantage of the two different forms of metadata. Section could be formally used to indicate the generic type of content in the entries and the key-value metadata used for specific information about the experiment. LabTrove has no mechanism for metadata hierarchies, but users could construct Key names to imply a hierarchy, for example, Material_Reagent, Material_Strain, and Material_Product. There are various ways that LabTrove metadata could be used to create taxonomies. Techniques usually applied to tags to create hierarchies could be applied to Keys, while key-value pair data could be extracted to create a hierarchical taxonomy, with the Keys representing one level and their values representing another.[16,49] Although this is a post hoc operation, it could be done on the fly, enabling the hierarchy to be displayed within the interface. Complex ontologies could be created from metadata items by making use of a "metadata operator" and chaining of multiple metadata values.[50] For example, a Key of "Conditions" could have multiple Values that themselves contain individual relationships, such as "Temperature:=200" and "Duration:=24 h", or even sequences of information, with appropriate delimiters defined, for example, "ID:=A1; Temperature:=200; Duration:=24 h; pH:=4.2".

4.2.2. Encouraging Addition of Metadata with Invitations and Profiles. Interface design has an important role in encouraging and supporting users to add their own metadata. User interfaces and application behavior can have a strong influence on the quality of metadata that users create.[6] Providing a visible and easy-to-use mechanism for adding metadata is essential. The benefits and functions of metadata cannot easily be exposed until metadata exists in a notebook, so the most important focus should be on enabling the user to add the metadata as simply as possible.

A very simple change that could be made to LabTrove would be to make the interface for adding metadata more inviting by using "action" words, for example, "Choose a Section", "Add a Key and Value", or "Create a new Key". A variety of techniques could be used to make the metadata creation facility more

prominent on the page, but an example that would not require a large redesign would be a Web 2.0 overlay that could be used to ask the user to "Add metadata to describe your experiment or entry" on creation of the entry or after the entry has been saved for the first time.

An alternative to asking the user to type content to be used as metadata is to allow the user to select previously used metadata values, as seen in the interfaces of Flickr, WordPress, Blogger, and myExperiment, or even to select existing content to use as a piece of metadata. The user could select a word or phrase and click a button to add the selected text (or image) as a tag. This selection of existing content to use as metadata is similar to the procedure in the A-Book ELN for Biologists, where the users draw a box around a name, procedure, drawing, or other object in order to label or categorize it for later search.[51]

An effective way to elicit quality metadata from the user is to make use of invitations in the form of questions. If we could ask the user specific questions about their experiment, then we could capture more information. Queries such as "What materials did you use in your experiment?" or "What instruments did you use in your experiment?" are focused on the vocabulary and viewpoint of the user and are therefore more likely to lead to the creation of relevant and meaningful metadata. Metadata that can still be stored using the existing metadata capabilities of LabTrove. Key-value pairs could be used for the storage of questions and answers, for example, using a Key of "Material" or "Instrument" with the associated values extracted from the input field. Multiple values could be captured in this way. Providing answers to question-based invitations would provide more consistency in the information generated and be more usable than the current interface. An "experiment profile" could be developed with standard questions that could be used in the interface to capture information relevant to an experiment. The high-level label-type values that represent a large proportion of the metadata used in the LabTrove community could be used as a starting point for this activity.

The community uses LabTrove for more than creating experiment records, so users could therefore be given the opportunity to select what kind of entry they are creating in their notebook. Each entry type, for example, project information, background literature, a plan, an experiment, or data, could have its own profile, with associated questions predefined for capturing relevant metadata. There is a risk that if no appropriate place is provided to record certain information, then that information will fail to be recorded,[52] so it is important to enable users to be able to define their own values if these are more appropriate. Communities and users should have the option to define their own entry types, such as sample description and instrument configurations that are used by some LabTrove communities, and appropriate properties for any type of entry. The Section could be used to store the information about what kind of entry the user has chosen to create.

In order to help determine what the most effective structures are for templates or profiles and what the best questions to ask in invitations might be, we are currently investigating how different structures and questions can affect what information researchers record for both experiments and metadata.[53,54]

4.2.3. Automatic Assistance. Alternatively, automatic methods could be used to prompt the user to add metadata in the same way that WordPress prompts users to add suggested tags based on the existing tags. The content of entries could be used to "guess" what the content may be about and prompt the user based on the results. For example, commonly used words can be extracted with data mining techniques such as word counts, matching words from a predefined vocabulary, identifying synonyms using a dictionary of terms, or creating metadata based on the format and context of the data files. Other examples include taking metadata from images by creating Keys and Values from the EXIF properties[55] and using a variety of tools to validate and extract the domain-specific metadata from data files.[56] Often data files are in tabular format, such as Excel files, or other structured formats, such as CSV files, where metadata could be extracted by examining the column or row headers. A similar function could be added to identify headers within notebook entries, for example, information formatted as a subheading or table headers.

Another example is to compare the notebook text to a dictionary of chemical terms while the user is writing the entry or at the point they save the entry. The user could then be prompted to add the chemical identifier as metadata or better still create the metadata automatically and then ask the user to verify it on saving the entry. The dictionaries used should contain any words relevant to the discipline, for example, equipment, instruments, or procedures. Individual users or teams could create their own custom dictionaries. These dictionaries could also contain mappings to URLs, so that all occurrences of the words could link to additional resources, for example, instructions and specifications for instruments or catalogues for chemicals.

Some work has already been done in LabTrove to perform automatic searching and matching of chemical names to create links to the relevant structures in the ChemSpider database.[57] This work could be extended to automatically create metadata for the entries based on the identified compounds.[58] For example, a key-value pair of "Material"-*Name of compound* could be created with the information derived from the data mining activity.

4.2.4. Visualizing Metadata. It is important to provide some immediate visual benefit from adding the metadata. For example, tag clouds could provide a very visual way to not only access the metadata and content beneath them but also to provide information about the use of the notebook. Not only tag-type metadata can be visualized using tag clouds, key-value metadata can be visualized by using the Key as the basis for the title and topic of the tag cloud. For example, a set of values for the Key "Materials" could be used to create a "Materials" tag cloud, while a "Topics" Key could create a "Topics" tag cloud. This information can be reinforcing for the authors of the notebook entries, encouraging them to add more categories and to associate entries with the appropriate metadata values. Tag clouds are also extremely valuable for helping others viewing the notebook to understand what it is about. The metadata can also be used to assist with filtering, grouping, and searching for information. Making these functions more visible, more powerful, and most importantly useful may increase the amount of metadata provided. Utilizing the interest that students have in social networking and the understanding that they have about metadata in that context may also provide alternative approaches for developing interfaces for metadata in an ELN for collaboration and open science.[24]

4.3. Educating Researchers. Our surveys show that Laboratory scientists need education, training, and encouragement to use metadata to curate their experiments and data.[16]

There are two different aspects of educating researchers that can be considered: (1) education about the metadata provisions available within LabTrove and (2) the vitally important need to provide researchers with education in the techniques and importance of metadata and data management in general.

4.3.1. LabTrove Education. Education is never a substitute for a usable interface, but education and training materials are still an important part of the user experience for software. The team have focused on significant improvements to the metadata documentation for LabTrove based on feedback from the LabTrove community. There is also the potential to add video tutorials and case study-style documentation to describe and demonstrate the use of the metadata provisions in LabTrove, appropriate workflows for adding metadata, and designing schemas for research groups. Courses for LabTrove could also be devised for students and researchers or be included as part of other laboratory and data management education. Exemplar notebooks that show best practices for metadata use and data management could also be made publicly available. Another possibility is to create "metadata checklists" that encompass recommended metadata items and a description for each item.[59] A different checklist could be created for different notebook uses and for different domains and could be made available to the wider research community.

4.3.2. Data Management Education. Despite the almost universal and regular use of technology by undergraduates,[60] there is evidence to show that the expectations that we have that researchers from the "Google Generation" will have high levels of skills and information literacy are unrealistic.[61] This observation is true even for researchers in the physical sciences, who are perceived to be more technical.[62] Research suggests that relatively few students actually know what metadata is or how to use it,[24] and knowledge of data management and curation is uncommon among researchers across all disciplines.[63−65] This suggests that one possible strategy for improving metadata use is to target users directly through data management education programs.[24,66−69]

At the University of Southampton, the majority of students receive some training in information retrieval at the beginning of their studies, usually delivered by library staff. These courses help them to develop skills at finding information from different sources, using filtering and Keywords, but further training is necessary to help them to make their own data "findable" in the future. As suggested by Swan and Brown,[66] there is a clear need to provide data management education to researchers at a postgraduate level, and it would be advantageous for undergraduates. Other studies have advocated the importance of appropriate data management training as early as possible in a researcher's career,[67] and we have previously advocated that instruction in data management should begin in school.[16] Such education should aid the student's understanding of organizing, classifying, and adding metadata to their own work, for their own benefit and for others, such as supervisors, the community, and the public.

The adoption of data management planning and education initiatives, such as the JISC Research data management (JISCMRD) projects such as the RDMP and RDMTrain projects, will contribute to improving the quality of metadata by providing examples of good practice and by developing data management training for researchers.[68,69] Education that is tailored to researchers and provides discipline and tool-specific training may be of more value to researchers than generic data management education.[67] The IDMB Archaeology Case Study provides an example of discipline-specific, practice-led, researcher training, including the introduction of a metadata model for use in a repository tool.[67]

Metadata use can be encouraged and facilitated by librarians and information specialists, who have data management expertise. Ideally, each research team would include a data management expert to provide support and advice and to help ensure best practice. If not, then an expert from outside could be brought in to provide assistance with the schema generation process and data management training[66] and maybe even "go native" to help.[16] Managers of projects could also take an active role by getting involved in the generation of schemas and using metadata quality indicators for assessing projects and data sets.[70]

5. CONCLUSIONS

The flexible metadata framework in LabTrove enables users to define metadata that is most appropriate for their work and experiments. The results of the LabTrove notebook survey demonstrate that simply providing the facility to add metadata is not enough to ensure that metadata is added to a notebook, let alone ensuring high quality metadata is added. For metadata to be useful it first has to be present, but enforcing metadata generation is of no benefit if it is low quality, inconsistent, or irrelevant. The LabTrove survey and the results of user research with the LabTrove community have indicated that the major inhibitors of metadata use in an ELN are the following:

- A lack of a defined metadata schema for a notebook or the "blank-canvas effect"
- A lack of knowledge about metadata
- The effort involved in the creation of metadata
- A lack of visibility and perceived benefits of metadata

The aim of the investigations into other platforms that enable users to create their own metadata was to determine whether the patterns of metadata usage by the LabTrove community were unusual and whether their interfaces might suggest enhancements that could be made to improve metadata capture within LabTrove.

The use of metadata in LabTrove may be lower than seen on the other platforms investigated, although it is difficult to tell because of the social nature of these platforms. The most comparable platforms to LabTrove in this regard are the NASA blogs, which shows a higher overall use of metadata but also show the same high use of a meaningless "catch-all" value.

The investigations indicated that the types of metadata that are added do vary across the different platforms and communities, but there are also many commonalities. In terms of word type, LabTrove is similar in verb and adjective use to both the NASA blogs and the Chemistry blogs. But myExperiment and Flickr use a greater proportion of verbs, and Flickr alone uses a significantly higher number of adjectives. LabTrove is notable for having a high percentage of activities described using noun-type words compared to the other platforms. This nominalization may reflect the research nature of LabTrove compared to the other platforms.[71] The subjects described in the metadata have commonalities across the platforms, such as materials, equipment, people, and locations. Each platform shows differences in the dominant subject types that are relevant to that platform. For example, in the NASA blogs, the dominant subject is equipment, describing entries about NASA-developed technologies such as rockets and satellite instruments, while myExperiment is dominated by

computer-related subjects describing software or data formats used in the experiment workflows. High-level labels dominate LabTrove Section values and Key terms, a pattern that is different from the metadata seen on the other platforms, although high-level terms such as "chemistry" and "education" are seen across all the platforms. Materials dominate the Value terms in LabTrove, and this term also makes up a high proportion of the Section terms and Key terms used, which is in common with the metadata from the Chemistry blogs investigated.

The metadata annotating digital documents must be present, of high quality, and consistent if it is to be useful. Given that a significant percentage of LabTrove users appear to have adopted a "minimum required" approach to metadata, working to encourage and support the creation of metadata is essential. Three different approaches could be used, either independently or in combination. The first approach is through the development of basic and community-appropriate metadata schemas and vocabularies, which can be used to seed LabTrove with metadata examples. Seeded metadata will help to mitigate the "blank canvas effect", by providing a starting point that users and teams can build upon, and can also improve quality by nurturing the use of consistent terms and abbreviations. The second approach is to apply data mining and other techniques to automate more capture of metadata by extracting key terms or guessing context by analyzing the content of the notebook entries. The final approach, and perhaps the most valuable in terms of encouraging the capture of user-defined metadata, is to enhance the user interface. Users can be encouraged to add metadata to their entries through the use of invitations prompting them to create metadata, reuse metadata, and provide information about specific elements of their experiment. More use can be made of the captured metadata within the interface, which in turn helps users to see the benefits of adding it and encouraging further metadata creation.

Our findings from these surveys have already had an influence on our ELN development activities in Southampton University and are likely to inform our ELN designs in the future.

■ ASSOCIATED CONTENT

Supporting Information

Numerical breakdowns and raw data of the results of the metadata surveys. Because many of the LabTrove notebooks included in the survey were private and not accessible to the public, the "raw" metadata from the LabTrove survey is not included. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail cerys.willoughby@soton.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research reported in this paper has been made possible by funding from the RCUK e-Science programme (EPSRC Grant GR/R67729, EP/C008863, EP/E502997, EP/G026238, BBSRC BB/D00652X), HEFCE and JISC Data Management Programme, and University Modernisation (UMF), and most recently the RCUK Digital Economy Theme as part of the IT as a Utility Network+ funding (EPSRC EP/K003569).

■ REFERENCES

(1) Boulton, G.; Campbell, P.; Collins, B.; Elias, P.; Hall, W.; Laurie, G.; O'Neill, O.; Rawlins, M.; Thornton, J.; Vallance P.; Walport, M. *Science as an Open Enterprise*, 2012. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf (accessed July 11, 2013).

(2) Ball, A. Metadata for Data Citation and Discovery. In *Disseminate, Discover: Metadata for Effective Data Citation*; DataCite, British Library, JISC Workshop, Opus: University of Bath Online Publication Store, 2012. http://opus.bath.ac.uk/30505/ (accessed July 11, 2013).

(3) Bird, C.; Willoughby, C.; Frey, J. Laboratory notebooks in the digital era: Record keeping in chemical and other science laboratories. *Chem. Soc. Rev.* **2013**, *42*, 8157−8175 DOI: 10.1039/c3cs60122f.

(4) Borgman, C. L. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT Press: Cambridge, MA, 2007; ISBN: 9780262514903.

(5) Borgman, C. L. Data, disciplines, and scholarly publishing. *Learn. Publ.* **2008**, *21*, 29−38.

(6) Crystal, A.; Greenberg, J. Usability of metadata creation application for resource authors. *Libr. Inf. Sci. Res.* **2005**, *27*, 177−189.

(7) Currier, S.; Barton, J.; O'Beirne, R.; Ryan, B. Quality assurance for digital learning object repositories: Issues for the metadata creation process. *ALT-J.* **2004**, *12*, 5−20.

(8) Ryan, B.; Walmsley, S. Implementing Metadata Collection: A Project's Problems and Solutions. *Learning Technol.* **2003**, *5*, Article 4. http://www.ieeetclt.org/issues/january2003/index.html#3 (accessed October 26, 2014).

(9) Pancerella, C.; Hewson, J.; Koegler, W.; Leahy, D.; Rahn, L.; Yang, C.; Myers, J. D.; Didier, B.; McCoy, R.; Schuchardt, K.; Stephan, E.; Windus, T.; Amin, K.; Bittner, S.; Lansing, C.; Minkoff, M.; Nijsure, S.; Laszewski, G.; Pinzon, R.; Ruscic, B.; Wagner, A.; Wang, B.; Pitz, W.; Ho, Y.; Montoya, D.; Xu, L.; Allison, T.; Green, W.; Frenklach, M. Metadata in the Collaboratory for Multi-Scale Chemical Science. In *DCMI '03 Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications: Supporting Communities of Discourse and Practice − Metadata Research & Applications*, September 28−October 2, 2003, Seattle, WA.http://dcpapers.dublincore.org/pubs/article/view/740 (accessed July 11, **2013**).

(10) Frey, J. Curation of Laboratory Experimental Data as Part of the Overall Data Lifecycle *Int. J. Digital Curation*, **2008**. *3*, 44−62. DOI: 10.2218/ijdc.v3i1. http://ijdc.net/index.php/ijdc/article/view/62/41 (accessed July 11, 2013).

(11) Milsted, A.; Hale, J.; Frey, J.; Neylon, C. LabTrove: A lightweight, Web-based, laboratory "blog" as a route towards a marked up record of work in a bioscience research laboratory. *PLoS One* **2013**, *8*, e67460 DOI: 10.1371/journal.pone.0067460.

(12) Frey, J. G.; Coles, S.; Bird, C. L.; Brocklesby, W. S.; Badiola, K. A.; Williamson, A. E.; Cronshaw, J. R.; Robins, M.; Roberston, M. N.; Todd, M. H.; Chapman, R. T.; Fisher, A.; Grossel, M.; Casson, J.; Hibbert, D. B.; Gloria, D.; Mapp, L. K.; Matthews, B.; Milsted, A.; Minns, R. S.; Mueller, K. T.; Parkinson, T.; Quinnell, R.; Robinson, J. S.; Springate, E.; Tizzard, G. J.; Willoughby, C.; Yang, E.; Ylioja, P. M.; Knight, N. J.; Marazzi, L. Experiences with a researcher-centric ELN. *Chem. Sci.* **2014**, DOI: 10.1039/C4SC02128B.

(13) Gilliland, A. J. Setting the Stage. In *Introduction to Metadata: Pathways to Digital Information*, version 3.0; Baca, M., Ed.; Getty Information Institute: Los Angeles, 2008. http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html (accessed October 17, 2014).

(14) Greenberg, J. A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *J. Assoc. Inf. Sci. Technol.* **2001**, *52*, 917−924 DOI: 10.1002/asi.1170.

(15) Zeng, M. L.; Qin, J. *Metadata*; Neal-Schuman Publishers: New York, 2008; ISBN: 9781555706357.

(16) Bird, C. L.; Willoughby, C.; Coles, S. J.; Frey, J. G. Data curation issues in the chemical sciences. *Inf. Stand. Q.* **2013**, *25* (3), 4−12 http://dx.doi.org/10.3789/isqv25no3.2013.02.
(17) Labtrove troves used in the study: http://altc.ourexperiment.org/, http://biolab.isis.rl.ac.uk, http://blogs.chem.soton.ac.uk/, http://www.ourexperiment.org/, http://xray.orc.soton.ac.uk/, https://www.enotebook.science.unsw.edu.au/, and http://mueller.ourexperiment.org.
(18) Attribute−Value Pair. Wikipedia.http://en.wikipedia.org/wiki/Key_Value_Pair (accessed May 14, 2014).
(19) Eynard, D.; Mazzola, L.; Dattolo, A. Exploiting tag similarities to discover synonyms and homonyms in folksonomies. *Softw: Pract. Exper.* **2012**, *12*, 1437−1457 DOI: 10.1002/spe.2150.
(20) Rahm, E.; Hai Do, H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* **2000**, *23* (4), 3−13.
(21) IBM. Generating a Taxonomy for Documents from Tag Data. U.S. Patent US 8346776 B2, 2013. http://www.google.com/patents/US8346776 (accessed July 31, 2013).
(22) Kjellberg, S. I am a blogging researcher: Motivations for blogging in a scholarly context. *First Monday* **2010**, *15*, Article 8. http://firstmonday.org/ojs/index.php/fm/article/view/2962/2580 (accessed July 31, 2013).
(23) Puschmann, C.; Mahrt, M. Scholarly Blogging: A New Form of Publishing or Science Journalism 2.0? In *Science and the Internet*; Tokar, A., Beurskens, M., Keuneke, S., Mahrt, M., Peters, I., Puschmann, C., Weller, K., van Treeck, T., Eds.; Düsseldorf University Press: Düsseldorf, 2012; pp 171−182.
(24) Mitchell, E. T. Metadata Literacy: An Analysis Of Metadata Awareness in College Students. Ph.D. Thesis, School of Information and Library Science, University of North Carolina at Chapel Hill, 2010. http://dc.lib.unc.edu/cdm/ref/collection/etd/id/2910 (accessed July 31, 2013).
(25) Ames, M.; Naaman, M. Why We Tag: Motivations for Annotation in Mobile and Online Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; Rosson, M. B., Gilmore, D., Eds.; Association for Computing Machinery: San Jose, CA, 2007; pp 971−980.
(26) Hammond, T.; Hannay, T.; Lund, B.; Scott, J. Social Bookmarking Tools (I): A General Review. *D-Lib Mag.* **2005**, *11*, article 1. http://www.dlib.org/dlib/april05/hammond/04hammond.html (accessed October 16, 2014).
(27) Flickr: The Commons. https://www.flickr.com/commons (accessed April 7, 2014).
(28) Discussing Machine Tags in Flickr API. https://www.flickr.com/groups/api/discuss/72157594497877875/ (accessed April 7, 2014).
(29) Proyecto Agua. Water Project on Flickr. https://www.flickr.com/people/microagua/ (accessed April 7, 2014).
(30) Flickr Hive Mind. http://flickrhivemind.net (accessed April 7, 2014).
(31) Most Interesting Pictures Tagged with Chemistry and Experiment from Flickr Hive Mind. http://flickrhivemind.net/Tags/chemistry,experiment/Interesting (accessed January 7, 2014).
(32) Yembrick, J. Personal communication, April 3, 2014.
(33) myExperiment. http://www.myexperiment.org (accessed April 7, 2014).
(34) De Roure, D.; Goble, C.; Stevens, R. *The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows*; Future Generation Computer Systems 25; University of Southampton: Southampton, U.K., 2009.
(35) All tags on myExperiment http://www.myexperiment.org/tags/ (accessed April 7, 2014).
(36) Sander, T.; Freyss, J.; von Korff, M.; Renée Reich, J.; Rufener, C. OSIRIS, an entirely in-house developed drug discovery informatics system. *J. Chem. Inf. Model.* **2009**, *49*, 232−246.
(37) Thibault, J. C.; Facelli, J. C.; Cheatham, T. E. iBIOMES: Managing and sharing biomolecular simulation data in a distributed environment. *J. Chem. Inf. Model* **2013**, *53*, 726−736 DOI: 10.1021/ci300524j.
(38) Facebook. https://www.facebook.com (accessed May 23, 2014).
(39) LinkedIn. https://www.linkedin.com (accessed May 23, 2014).
(40) Summary of Peer Review Harmonisation Activities. Research Councils UK. http://www.rcuk.ac.uk/research/efficiency/researchadmin/harmonisation/ (accessed July 31, 2013).
(41) RCUK Research Classifications (data file). http://www.rcuk.ac.uk/research/Efficiency/Pages/harmonisation.aspx (accessed November 2014).
(42) Coles, S. J.; Frey, J. G.; Bird, C. L.; Whitby, R. J.; Day, A. E. First steps towards semantic descriptions of electronic laboratory notebook records. *J. Cheminf.* **2013**, *5:52*, 1−10 DOI: 10.1186/1758-2946-5-52.
(43) Borkum, M.; Frey, J. G.; Coles, S. oreChem: Planning and Enacting Chemistry on the Semantic Web. Microsoft Research eScience Workshop 2010, Berkeley, CA, October 11−13, 2010. http://www.slideshare.net/mark.borkum/orechem-planning-and-enacting-chemistry-on-the-semantic-web (accessed October 22, 2014).
(44) Lagoze, C. The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web. In *Proceedings of the WebSci'09: Society On-Line*, Athens, Greece, 2009.
(45) Kaestle, G.; Shek, E. C.; Dao, S. K. Sharing Experiences from Scientific Experiments. In *Proceedings of the 11th International Conference on Scientific on Scientific and Statistical Database Management*, Cleveland, OH, July 28−30, 1999, pp 168−177.
(46) Sufi, S.; Matthews, B.; Kleese van Dam, K. An interdisciplinary model for the representation of scientific studies and associated data holdings. UK e-Science All Hands Meeting. Nottingham, U.K., September 2−4, 2003, pp 103−110. http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/020.pdf (accessed November 2104).
(47) Wang, F.; Liu, P.; Pearson, J.; Azar, F.; Madlmayr, G. Experiment Management with Metadata-based Integration for Collaborative Scientific Research. ICDE'06 Proceedings of the 22nd International Conference on Data Engineering, IEEE Computer Society, Atlanta, GA, April 3−7, 2006, p 96. DOI: 10.1109/ICDE.2006.65.
(48) Chemical Methods Ontology (CMO). Royal Society of Chemistry. http://www.rsc.org/ontologies/CMO/index.asp (accessed October 22, 2014).
(49) Heymann, P.; Garcia-Molina, H. *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*; Technical Report 2006-10; Stanford University: Stanford, CA, 2006. http://ilpubs.stanford.edu:8090/775/1/2006-10.pdf (accessed July 31, 2013).
(50) Robson, B. Clinical and pharmacogenomic data mining: 4. The FANO program and command set as an example of tools for biomedical discovery and evidence-based medicine. *J. Proteome Res.* **2008**, *7*, 3922−3947 DOI: 10.1021/pr800204f.
(51) Mackay, W. E.; Pothier, G. The A-Book: An augmented laboratory notebook for biologists. *ERCIM News* **2001**, *46*, 52−53.
(52) Swinglehurst, D.; Greenhalgh, T.; Roberts, C. Computer templates in chronic disease management: Ethnographic case study in general practice. *BMJ Open* **2012**, *2*, e001754 DOI: 10.1136/bmjopen-2012-001754.
(53) Willoughby, C.; Bird, C.; Frey, J. User-defined metadata: Using cues and changing perspectives. 2014, submitted.
(54) Willoughby, C.; Frey, J. Effects of using structured templates for recording chemistry experiments. 2014, in preparation.
(55) Exchangeable Image File Format. Wikipedia. http://en.wikipedia.org/wiki/Exchangeable_image_file_format (accessed May 23, 2014).
(56) Downing, J.; Murray-Rust, P.; Tonge, A. P.; Morgan, P.; Rzepa, H. S.; Cotterill, F.; Harvey, M. J. SPECTRa: The deposition and validation of primary chemistry research data in digital repositories. *J. Chem. Inf. Model.* **2008**, *48*, 1571−1581 DOI: 10.1021/ci7004737.
(57) Chemspider. http://www.chemspider.com/About.aspx? (accessed July 29, 2013).
(58) Day, A. Search and Insert from ChemSpider in LabTrove and Other Websites with TinyMCE Editors, 2013. http://www.chemspider.com/blog/search-and-add-from-chemspiderdirectly-from-

labtrove-and-other-blog-based-websites-which-use-tinymce.html (accessed July 11, 2013).
(59) Higdon, R.; Stewart, E.; Stanberry, L.; Haynes, W.; Choiniere, J.; Montague, E.; Kolker, E. MOPED enables discoveries through consistently processed proteomics data. *J. Proteome Res.* **2014**, *13*, 107−113 DOI: 10.1021/pr400884c.
(60) Smith, S. D.; Salaway, G.; Caruso, J. B. The ECAR Study of Undergraduate Students and Information Technology. EDUCAUSE Center for Applied Research, 2009. http://net.educause.edu/ir/library/pdf/ers0906/rs/ers0906w.pdf (accessed July 31, 2013).
(61) Rowlands, I.; Nicholas, D.; Williams, P.; Huntington, P.; Fieldhouse, M.; Gunter, B.; Withey, R.; Jamali, H. R.; Dobrowolski, T.; Tenopir, C. The Google generation: The information behaviour of the researcher of the future. *Aslib Proc.* **2008**, *60*, 290−310.
(62) Goodger, J.; Worthington, W. Research Data Management Training for the Whole Project Lifecycle. In *Physics & Astronomy Research (RDMTPA) JISC Final Report*. Research Data Toolkit, University of Hertfordshire, 2013. http://research-data-toolkit.herts.ac.uk/document/rdmtpa-final-report/ (accessed July 11, 2013).
(63) Whitton, M.; Takeda, K. Data Management Questionnaire Results: IDMB Project. University of Southampton: Southampton, GB, 2011. http://eprints.soton.ac.uk/196243/1/IDMB_Survey_Report.pdf (accessed July 11, 2013).
(64) Whitton, M. Institutional Data Management Blueprint (IDMB) Questionnaire Quantitative Data. University of Southampton: Southampton, GB, 2011. http://eprints.soton.ac.uk/195959/ (accessed July 11, 2013).
(65) Ward, C.; Freiman, L.; Jones, S.; Molloy, L.; Snow, K. Making Sense: Talking Data Management with Researchers. *Int. J. Digital Curation*. **2011**, *6*, 265−273. http://www.ijdc.net/index.php/ijdc/article/view/197 (accessed July 11, 2013).
(66) Swan, A.; Brown, S. Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs; Report to the JISC; 2008. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf (accessed July 11, 2013).
(67) Earl, G.; White, W.; Wake, P. *IDMB Archaeology Case Study: Summary*; University of Southampton: Southampton, GB, 2011. http://eprints.soton.ac.uk/id/eprint/196237 (accessed July 11, 2013).
(68) Managing Research Data Programme 2011-13. JISC. http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx (accessed July 11, 2013).
(69) Hodson, S. Meeting the Research Data Challenge JISC Briefing Paper, 2009. http://www.jisc.ac.uk/publications/briefingpapers/2009/bpresearchdatachallenge.aspx#downloads (accessed July 11, 2003).
(70) Bruce, T. R.; Hillmann, D. Metadata Quality in a Linked Data Context. *VoxPopuLII*, 2013. http://blog.law.cornell.edu/voxpop/2013/01/24/metadata-quality-in-a-linked-data-context/ (accessed July 31, 2013).
(71) Biber, D.; Gray, B. Nominalizing the Verb Phrase in Academic Science Writing. In *The Verb Phrase in English*, 1st ed.; Aarts, B.; Close, J.; Leech, G.; Wallis, S., Eds.; Cambridge University Press: Cambridge, 2013; pp 99−132. http://dx.doi.org/10.1017/CBO9781139060998.006 (accessed November 2014).