Creating context for the experiment record: user-defined metadata

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

Chemistry, University of Southampton

Introduction

The drive towards more transparency in research and open data increases the importance of being able to find information and make links to the data. Metadata is an essential part of this process and for the preservation of knowledge for future exploitation. Metadata is often defined as "data about data" but can better be defined as the information that describes a physical or digital document or object[1]. Metadata provides context to the data and enables relationships between different data to be explored, making the data more usable and reusable, persistent, discoverable, and accessible[2].

Metadata is used in Electronic Laboratory Notebooks to curate experiment data and associated entries with descriptive information and classification labels that can be used for aggregation and identification. Machine-generated metadata helps with facilitating metadata exchange and enabling interoperability, but such metadata is not necessarily in a form friendly for the humans that also need it.

Metadata in LabTrove

LabTrove, a researcher-centric web- and cloud-based ELN developed at the University of Southampton, enables users to add their own user-defined metadata. LabTrove has a blog-style structure that enables users of the system to record their experiments and activities with individual entries in the notebook. Each entry must have a user-defined value for "section" and users can optionally choose to add further metadata to their entries in the form of "key-value pairs". A key-value pair is a data representation that is used to represent a unique property that can have many different associated values. The key-value pairs enable the inclusion of metadata that is much richer than could be produced using a simple tagging system: key-value pairs provide a form of classification for notebook entries. The use of consistent metadata potentially produces a much more effective record than a paper notebook[3].

User-defined metadata surveys

We surveyed 104 LabTrove blogs from a variety of users across the globe to investigate patterns of metadata usage to identify whether metadata was being used effectively, potential strategies for encouraging metadata use, and ways in which the user experience might be improved. The findings of the survey indicated that many of our users were not using metadata effectively. Figure 1 shows the breakdown of "section" and "key" metadata elements used for each notebook, highlighting that relatively high numbers of users are using a minimum amount of metadata, including 1/3 that use only one section, and 50% of users using an unhelpful "catch-all" section. Few users use large numbers of metadata elements to describe their notebook entries, although the number of elements does increase with the number of authors working on a single notebook. The survey information coupled with information from interviewing users and conducting user studies indicates that, whilst some groups are comfortable with metadata and are able to design a metadata structure that works effectively, many users have no knowledge of where to start to define metadata or even an understanding of what it is and why it is useful. We also found that the metadata used within the notebooks is dominated by a few categories, in particular high-level labels, and elements describing materials, data formats, and instruments. One of the observations of the study was that the metadata used in the notebooks was primarily about "things" rather than "activities" with little use of verbs and adjectives.

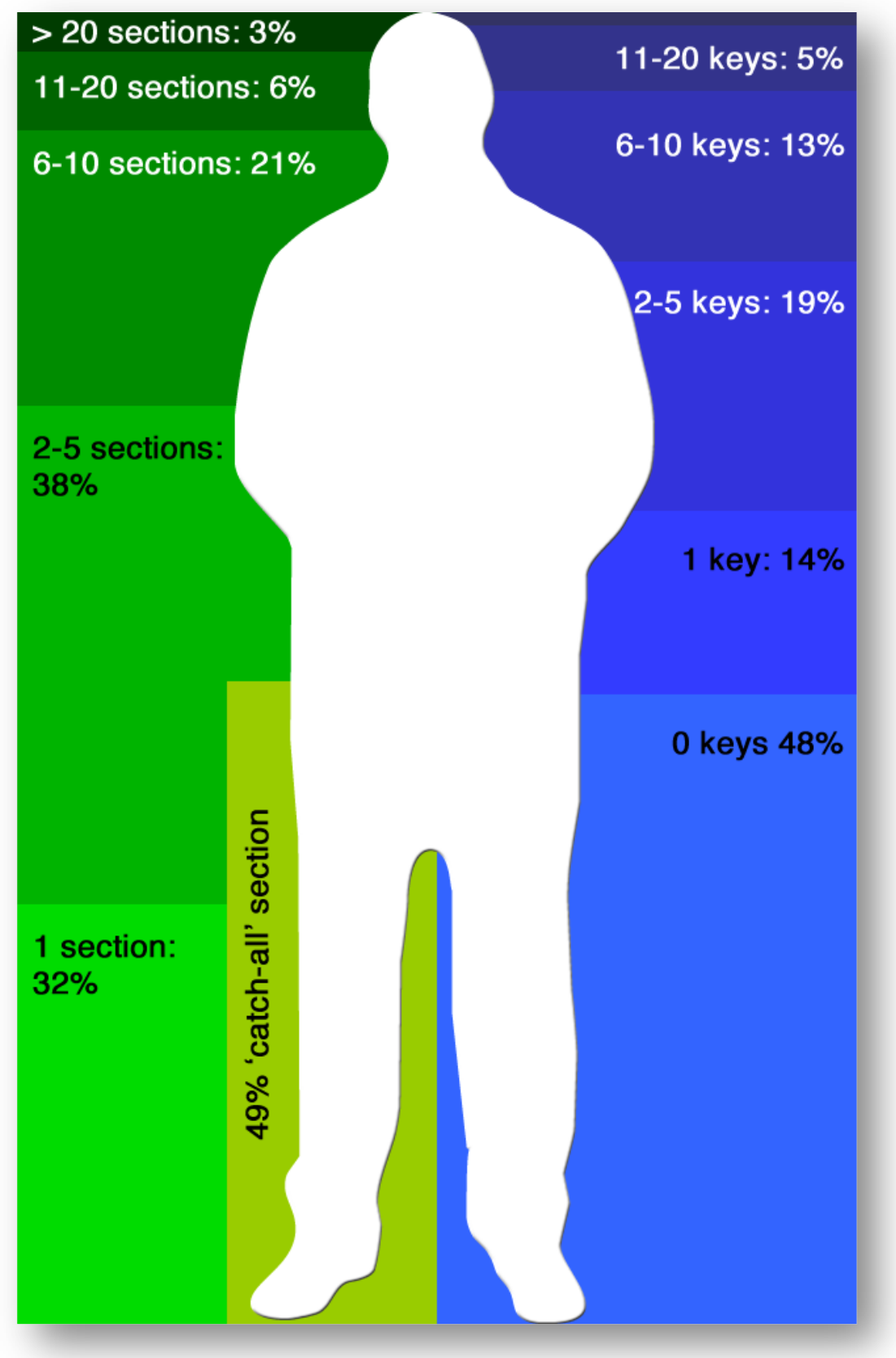
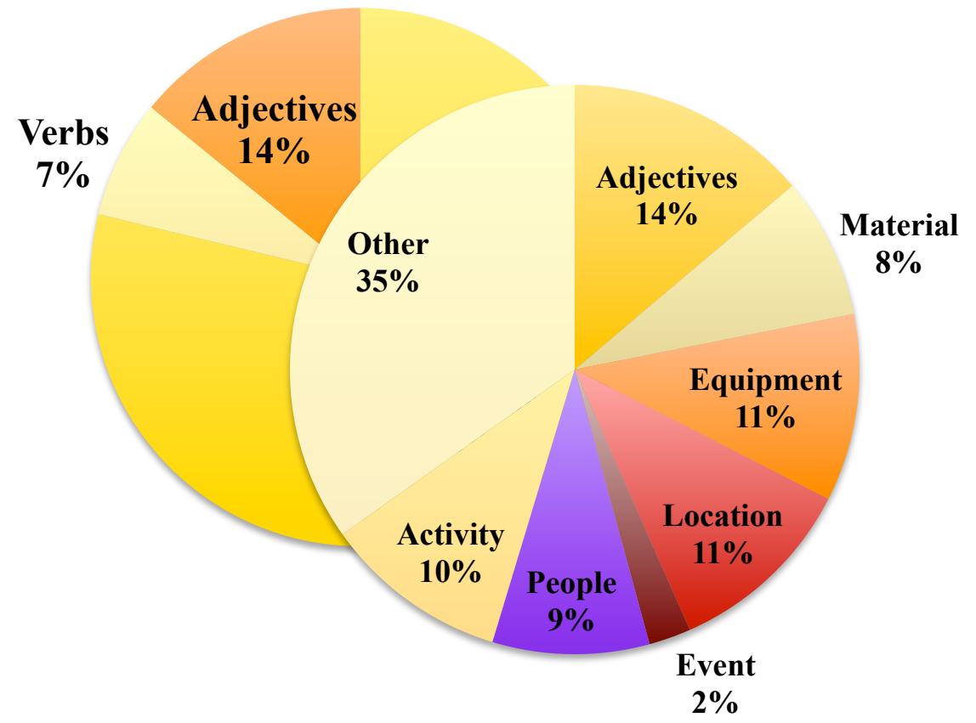


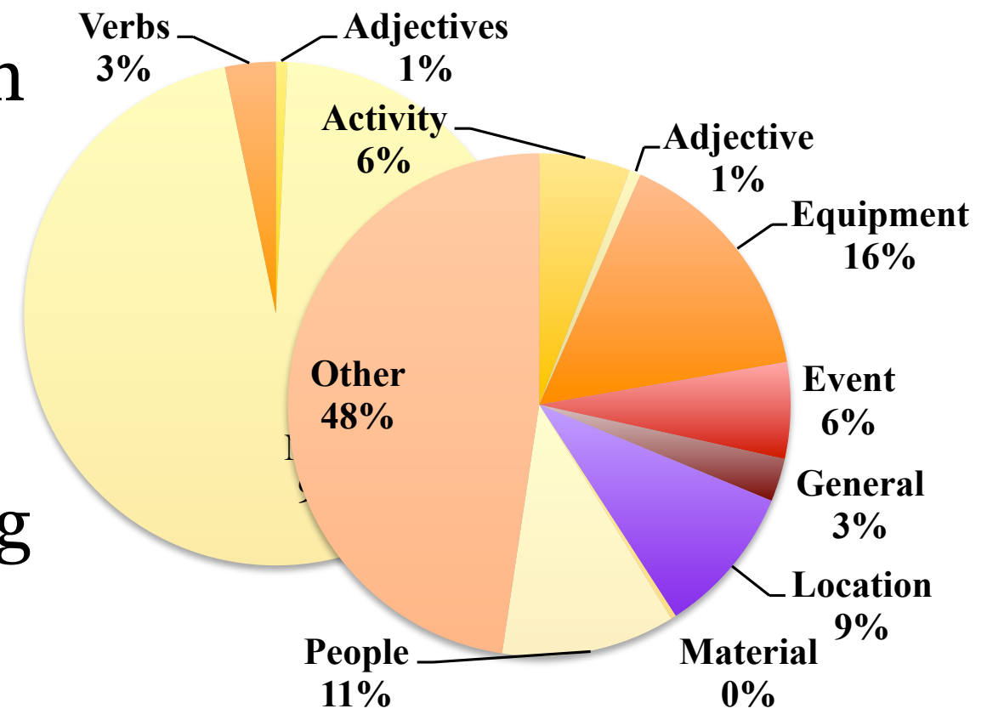
Figure 1: Numbers of sections and keys used in each notebook

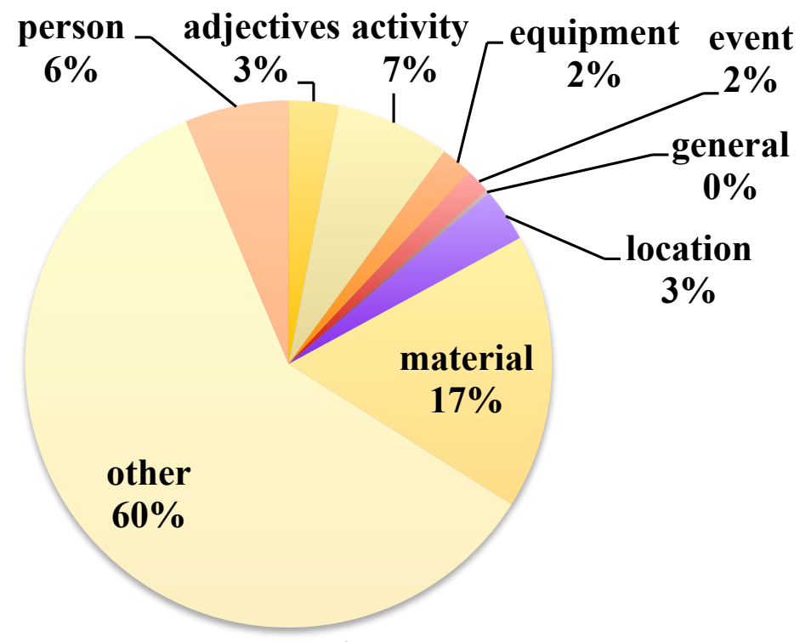
Further investigation was carried out to determine whether the pattern of metadata use we observed in LabTrove was common in other online environments where users could add their own metadata. The aim of the surveys was to determine whether users are more likely to create certain types of metadata and whether lessons can be learned from other environments to encourage metadata use. Metadata elements from each environment were categorized by word type "noun", "verb", and "adjective" (the noun category includes all words, phrases, and abbreviations that represent an object). Each noun was further categorized into groups relating to "materials", "equipment", "locations", "events", "people", "activities", "other", and "general" where a catch-all type of metadata value is used.



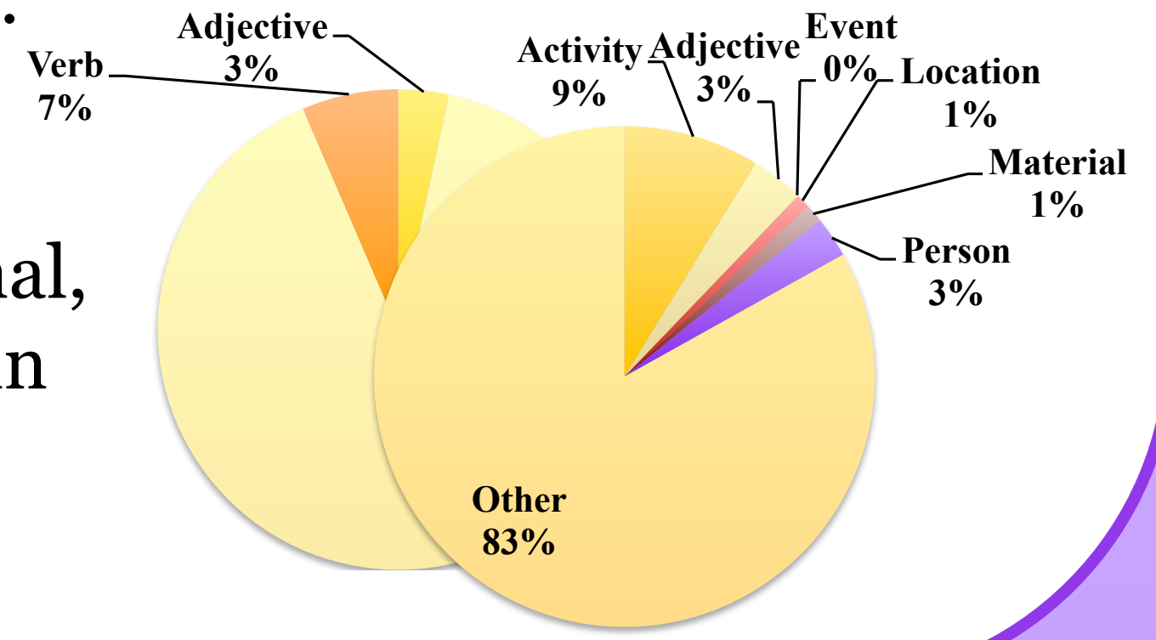
Flickr was surveyed as an environment where user-defined metadata in the form of tags is well-used by the community. 1381 unique tags associated with 500 photos tagged with "chemistry" and "experiment" were categorized for the survey. The results showed that adjectives are used more often in this environment than in the other environments surveyed, and that none of the groups is particularly dominant in the noun-type category.

1226 metadata elements from 53 NASA blogs were surveyed as an open environment with a similar diversity to LabTrove with a variety of blog topics, between 1 and 45 authors, and a minimum amount of metadata that can be used. Only 17% of the blogs use a single element, with 3 blogs using well over 100 elements. The results showed that verbs and adjectives are used even less in this environment than in LabTrove. "Equipment" is the dominant group in the noun-type category, together with location and people highlighting the many NASA missions and facilities, and their interactions with other communities.





10,436 metadata elements from 50 chemistry-related blogs, mostly based on the WordPress and Blogger platforms, were surveyed. The results showed that nouns made up 94% of the metadata used, and that "materials", such as chemicals, biological samples, and medicinal drugs, are the dominant group. Other groups in the noun category were evenly represented, with very little use of catch-all metadata.



2349 unique tags were surveyed from the workflows on myExperiment, chosen as a community that shares activities through scientific workflows. Adding tags is optional, but but the vast majority of users use between 1 and 10 tags. The noun groups seen in the other surveyed environments are poorly represented in myExperiment, whilst computing-related software, topics, and abbreviations are the most dominant type.

Futures

The findings from our LabTrove metadata study has already been used to help influence the design of our most recent mobile ELN and the companion experiment plan tool. Interfaces were designed to capture useful metadata by prompting the users to enter information about their experiments using the headings identified from the "labels" present in the survey and user research with ELN users. These experiment records and associated metadata can be exported into LabTrove.

We are also investigating alternatives for experiment markup and whether providing cues changes the metadata that is recorded.

Conclusions

Metadata has to be present to be useful, and interface designs can encourage more effective metadata use, such as visibility, viewing previously used values, autocorrect, suggestions, missing metadata, and providing meaningful defaults. Metadata used tells us about the interests and needs of communities, and can provide a basis for formal taxonomies and markup valuable for those communities.

Literature

1. Zeng, M. and Qin, J. *Metadata* Neal-Schuman: New York. 2008 ISBN: 978-1555706357

2. Kowalczyk, S. and Shankar, K. Data sharing in the sciences. *Ann. Rev. Info. Sci. Tech.* 2011, 45: 247–294. doi: 10.1002/aris.2011.1440450113

3. Bird, C. Willoughby, C., and Frey, J. Laboratory notebooks in the digital era: Record keeping in chemical and other science laboratories *Chem. Soc. Rev.*, 2013,42, 8157-8175

Acknowledgements

Smart Research Framework, e-Research South. Emma Tonkin and postgraduate students at the University of Bath. Organic Synthesis students and LabTrove users at the University of Southampton. LabTrove users at University of Sydney, University of New South Wales, and Penn State University.