Identifying Description Indicators for Research Data from Scientific Journal Publications

Tiffany C. Chao[1]

[1] Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Abstract

In order to support the sharing and reuse of scientific research data, rich description about the data must be made available. Scientific journal publications are a potential resource in contributing contextual details about the collection, generation, use, and analysis of data critical for facilitating meaningful interpretation. This poster presents an exploratory study on what information related to data can be identified from published literature on soil science research. The preliminary findings reveal the range of information detailed about data within journal publication including discussion of data sources, referenced techniques and processes applied to data, and description on how data variables were collected and derived. With the growth of digital data, these findings will contribute to the development of a systematic approach for enhancing description in data curation systems and services and fostering data reuse.

Keywords: research data description, data curation, data reuse, scholarly publication

Citation: Chao, T. C. (2014). Identifying Description Indicators for Research Data from Scientific Journal Publications. In *iConference 2014 Proceedings* (p. 1038–1042). doi:10.9776/14366

Contact: tchao@illinois.edu

1 Introduction

The provision of description and metadata are essential for the discovery, sharing, and reuse of research data. However, obtaining such information from those involved in producing the data is a time- and resource-intensive process. Descriptions are beneficial for accounting what data have been collected and are available, but can also provide insight on how and why data were created, and explain anomalies or areas of uncertainty that arose during the research process. The current emergence of scientific workflow system adoption demonstrates an automated alternative to manually documenting data production throughout the research lifecycle (Littauer et al., 2012). Other systems, such as the UniProt (http://www.uniprot.org/) database for protein data, also curate annotations both automatically and manually generated which contribute to a more robust provenance record for the data. However, the use of scientific workflows or automated tools for documentation is still not widespread across scientific domains including small science research (Davis et al., 2012). Small science research studies garner a significant portion of scientific funding in the US yet the ad hoc documentation and use of metadata standards make the data generated from these studies difficult to readily access or reuse by others (Heidorn, 2008; Wallis, Rolando, & Borgman, 2013). Similarly, survey findings reported by Tenopir et al. (2011) suggest scientists generally are not active in applying metadata to describe their datasets with only some who utilize locally developed standards. These contrasts and variations in metadata use and documentation practice for data further exacerbates the challenge of securing description information to foster future use of the data. With increased attention to the development of infrastructure and services for the curation and long-term management of research data in libraries, archives, and repositories, identifying an approach to procure description information for available data is needed.

Data are a key part of the foundation underlying scholarly journal publications and increasingly becoming accessible as supplements to published articles (Borgman, 2012) or embedded as part of online

journal publications allowing for user interaction and annotation (Attwood et al., 2010; Renear & Palmer, 2009). Scientific journals publications remain a primary mechanism of communication among scientists and scholars, advancing scientific knowledge and innovation and providing meaningful information units for further discussion and analysis (Brown, 2010). The descriptive content and embedded data representations (e.g. figures, tables, charts, etc.) of journal articles also play a vital role for researchers to verify the reliability of data for reuse (Faniel & Jacobsen, 2010) or as information sources to discover data for new inquiry (Davis et al., 2012). Given the prominent role of journal articles within the scientific community for communicating scholarly information and as a resource for data discovery and study, there is potential for publications to be used as a source for informing data description for curation. This study investigates what indicators related to data can be identified within the content of journal publications to support continued curation of research data.

2 Method

In this exploratory study, nine full-text articles were collected from three peer-reviewed journals in the soil sciences: Soil Science Society of America Journal, Applied Soil Ecology, and European Journal of Soil Biology. The selected journals are considered top tier in the field based on published rankings from Scimago (http://www.scimagojr.com/) and Thomson Reuters Journal Citation Reports. Soil science is investigated as it is representative of small science research where data generated are in high need of curation support and primarily analyzed and used locally within a research group (Cragin et al., 2010). In addition, the rigorous research data collection procedures and generation of heterogeneous data types for analysis, along with the rise in meta-analysis research which necessitates consultation of different datasets and results, suggests that a high level of detail related to the data will be documented and represented within soil science publication content.

As a starting point, research articles published between 2006-2011 were selected at random for this exploratory sample. Descriptive coding of the articles was manually performed, with the initial codelist derived from a functional vocabulary introduced by Cragin, Palmer, and Chao (2010) that maps relationships between data characteristics, research practices, and curation activities, and gradually refined with emerging themes based on subsequent rounds of coding.

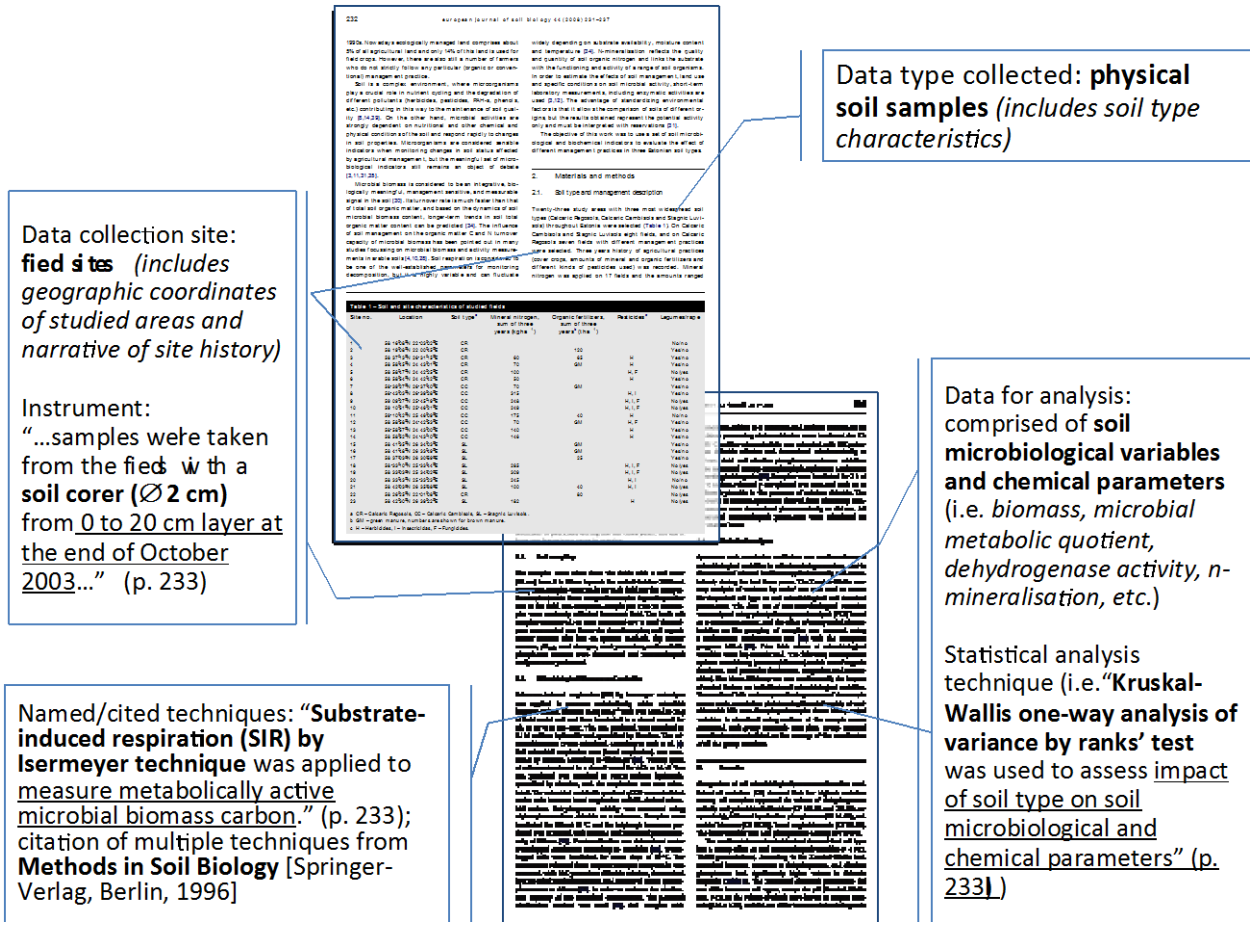

Figure 1: Examples of indicators identified from sample article (Truu, Truu, & Ivask, 2008): includes how these data details are represented in-situ to describe the context of data collection and processing for analysis.

3 Preliminary Findings

Across all the articles from the different journals, several themes became visible regarding available information related to data and associated research practices. Figure 1 details an example of the description indicators identified from a journal article on soil microbiological and biochemical properties assessment. The article encompasses details of the study site where collection of soil samples occurred (data collection site; data type collected), the instruments and techniques applied in collecting and processing the soil samples including units of measurement (instrument; named/cited technique), and what soil microbiological variables comprise the dataset for statistical analysis (statistical analysis technique; data for analysis).

In considering curation implications based on the available indicator details, the description of the study site provides rich contextual evidence regarding the data source which contributes to the provenance of the data. The applied techniques and instruments used are critical for future replication and may also provide insight to known standards within a given discipline for techniques that are used to generate particular data. These description indicators of *data collection site, data type collected, instruments, named or cited techniques, data for analysis,* and *statistical analysis technique* were consistently observed in the sample, although the degree of elaboration for each indicator varied between articles within the same journal. Some articles also detailed quality control practices, such as the removal of particulates that surpassed a certain threshold and homogenization of soil samples for analysis. Additional sources of description information about data were found in the succinct captions for tables and figures, often relaying

resulting relationships between different data variables; aligning what variables were assessed or measured with how they were generated or derived appears to be possible based on the details of a publication and may contribute to the value of these data.

4 Future Work

To maximize the potential for data sharing and reuse, the provision of rich data description is necessary. Preliminary results from this study propose journal publications are a productive resource for distinguishing contextual information about the collection of data and how these details are represented, such as cited references for techniques used or numerical values. Next steps include increasing the publication sample size within soil science for more detailed analysis of journal article content to solidify observed themes. Specific attention will be given to trends in cited references for techniques and developing a more systematic approach to determining the presence of a data description indicator. It will also be helpful to see how this approach extends to other disciplines within the small sciences. Additional exploration of available tools is needed to more fully understand how description information for data can be extracted from research articles. Establishing a concrete base for indicators can have potential implications for the development and advancement of automated processes to capture and enhance data description in supporting data repositories and curation services.

5 References

Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., & Thorne, D. (2010). Utopia documents: linking scholarly literature with research data. *Bioinformatics*, *26*(18), i568-i574.doi:10.1093/bioinformatics/btq383

Borgman, C. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634

Brown, C. (2010). Communication in the sciences. *Annual review of information science and technology*, *44*(1), 285–316.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. doi:10.1098/rsta.2010.0165

Cragin, M. H., Palmer, C. L., & Chao, T. C. (2010). Relating data practices, types, and curation functions: An empirically derived framework. *Proceedings of the American Society for Information Scienceand Technology*, *47*(1), 1–2.

Davis, L., Qin, H., D'Ignazio, J., Romero Lankao, P., Mayernik, M., & Alston, P. (2012). *Variables as currency: linking meta-analysis research and data paths in science.* [White paper]. Retrieved from http://dlsciences.org/research/DataConservancy/Variables%20as%20Currency.pdf

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(34), 355–375. doi:10.1007/s10606-010-9117-8

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299.

Littauer, R., Ram, K., Ludäscher, B., Michener, W., & Koskela, R. (2012). Trends in use of scientific workflows: insights from a public repository and recommendations for best practice. *International Journal of Digital Curation*, *7*(2), 92–100. doi:10.2218/ijdc.v7i2.232

Renear, A. H., & Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science*, *325*(5942), 828–832. doi:10.1126/science.1157784

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PloS one*, *6*(6), e21101.

Truu, M., Truu, J., & Ivask, M. (2008). Soil microbiological and biochemical properties for assessing the effect of agricultural management practices in Estonian cultivated soils. *European Journal of Soil Biology*, *44*(2), 231–237. doi:10.1016/j.ejsobi.2007.12.003

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), e67332. doi:10.1371/journal.pone.0067332

6 Table of Figures