Exploring description for research data in soil science journal publications

Tiffany Chao (tchao @ illinois.edu)
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign

OVERVIEW

An essential component to sharing research data with others is the provision of metadata to facilitate meaningful interpretation. Speaking with the original data producer is one of the few approaches available for curators to obtain metadata yet involves substantive time and resource investment to coordinate and conduct. In order to optimize this time with data producers, the objective of this pilot study is to understand what information contained in journal publications by data producers can be used to inform metadata description for data.

A secondary aim is to explore how this extracted description for data can be applied to support current curation services in libraries and institutional repositories. The Data Curation Profile[1] (Profile) was chosen as a framework to guide identification of potential metadata. The Profile is an established tool developed to capture researcher expectations and requirements for the curation of their research data with specific sections for describing data kinds and the different stages that correspond to the research process and lifecycle.[2]

METHOD

A sample of (15) articles was selected from three peer-reviewed journals in the soil sciences published between 2006-2012. The discipline of soil science is representative of small science research where data are characterized as heterogeneous in format with ad hoc application of data standards and deemed in high need of curation support.[3] The selection of journals for the sample—Soil Science Society of America Journal, Soil Biology and Biochemistry, and Plant and Soil—comprise a variety of publishers within the subject area which can contribute to differences in descriptive information related to data identified from journal articles.

Articles were manually annotated using the Profile sections (see Table 1) and subsections for guidance. Particular attention was given to the following sections: *Overview of research* (Sec. 2), *Data kinds and stages* (Sec. 3), and *Organization and description of data* (Sec. 5). For this pilot study, each article results in an individual Profile.

Table 1. Data Curation Profile Sections

Section 2 - **Overview of the research**	Section 8 - **Discovery**
Section 3 - **Data kinds and stages**	Section 9 - **Tools**
Section 4 - **Intellectual property context and information**	Section 10 - **Linking / Interoperability**
Section 5 - **Organization and description of data (incl. metadata)**	Section 11 - **Measuring Impact**
Section 6 - **Ingest / Transfer**	Section 12 – **Data Management**
Section 7 - **Sharing & Access**	Section 13 - **Preservation**

JOURNAL ARTICLE ANALYSIS EXAMPLE

Selected annotations for sample article, Alvaro-Fuentes et al., (2009).[4]

(p.1520) "Consequently, the overall objective of the present work was to determine the effects of different tillage and cropping systems on soil C stabilization by soil aggregates in Mediterranean semi- arid conditions… In the first experiment (Exp. 1), soil C fractions were isolated from size-class aggregates in different tillage and cropping systems. The second experiment (Exp. 2) was set up to investigate the role of microaggregates occluded within macroaggregates in the long-term SOC sequestration in Mediterranean semiarid conditions."

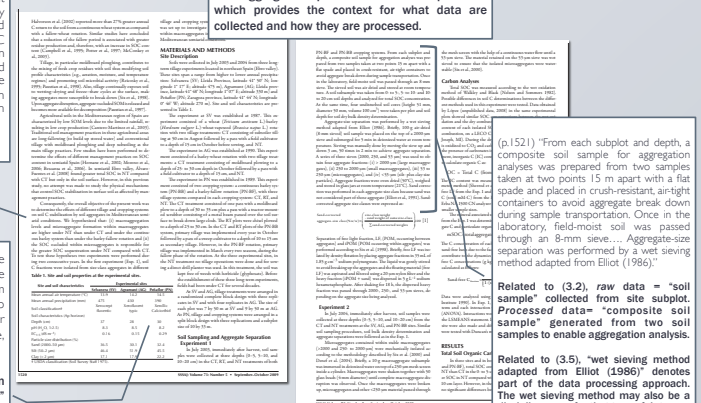
Related to (2.1), describes research goals of study

(p.1520) "Table 1. Site and soil properties at the experimental sites." "Site Description. Soils were collected in July 2003 and 2004 from three long- term tillage experiments located in northeast Spain (Ebro valley). These sites span a range from higher to lower annual precipitation: Selvanera (SV; Lleida Province, latitude 41° 5′ N; longitude 1° 17′ E; altitude 475 m)"

Related to (3.5), describes site of data collection (incl. latitude and longitude). "Table 1" complements the site description narrative.

(p.1520-21) "Soil Sampling and Aggregate Separation." "Experiment 1" and "Experiment 2"

Article sections related to (3.5), the soil sampling and aggregate separation for each experiment which provides the context for what data are collected and how they are processed.



(p.1521) "From each subplot and depth, a composite soil sample for aggregation analysis was prepared from two samples taken at two points 15 m apart with a flat spade and placed in crush-resistant, air-tight containers to avoid aggregate break down during sample transportation. Once in the laboratory, field-moist soil was passed through an 8-mm sieve…. Aggregate-size separation was performed by a wet sieving method adapted from Elliot (1986)."

Related to (3.2), *raw* data = "soil sample" collected from site subplot. *Processed* data= "composite soil sample" generated from two soil samples to enable aggregation analysis.

Related to (3.5), "wet sieving method adapted from Elliot (1986)" denotes part of the data processing approach. The wet sieving method may also be a discipline norm for that type of data.

Observations of Data Curation Profile utility across all articles in sample

Section 2:
2.1- Research area focus
- Often stated in Abstract or as part of the background sections of an article

2.3- Funding sources
- Noted in the 'Acknowledgments' section of an article

Section 3:
3.1- Data narrative
- Parts of the Narrative can be derived from sections of the article but not all stages of the data lifecycle can be appropriately accounted for

3.2- Data table (including "data stages": raw, processed, analyzed, finalized)
- Similar to the Narrative, there are gaps in completing the Table

3.5- Contextual narrative
- The majority of descriptive information about the data seemed to fit in this section

Section 5:
5.2- Formal standards used
5.3- Locally developed standards
- While these standards are specific to the organization of data, the high frequency of referenced procedures provides insight to potential 'standards' for methods and practice within the soil science community

PRELIMINARY FINDINGS

Completing a Profile requires additional information not found in journal publications, but certain sections may be more readily answered based on information provided in these articles. There did not appear to be significant differences in findings across the different journals.

Three research practices are consistently described across all journal articles:
- sampling procedures for gathering data (i.e. physical samples)
- processing physical samples using particular instrumentation and procedures
- conducting statistical analysis on the processed data

These practices can potentially indicate information related to the data stages of the Profile. For instance, "sampling" is closely connected with description of *raw* data while "processing" describes *processed* data. It can be inferred that the statistical analysis of processed data would result in *analyzed* data.

Other sections of the Profile that could be recorded based on journal article content include *Tools* (Section 9), though not as frequent.

Some data kinds, such as "soil organic carbon" appear more regularly across the Profiles with variations in the research practices used for sampling and processing.

FUTURE WORK

The journal articles from the soil sciences provide insight to the processes and practices related to how data emerge and have potential use for imparting descriptive metadata for data that can contribute to curation efforts. As journal articles continue to be put forth by data producers, more effective approaches are needed to harness content about data from publications.

The initial findings provide the basis for additional areas of exploration, which include:
- Examining the relationship between multiple articles generated from a single dataset
- Developing a framework to more systematically identify information from journal articles that can be used as metadata

References: [1]Data Curation Profiles (http://datacurationprofiles.org/); [2]Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, 4(3). 93–103. doi:10.2218/ijdc.v4i3.117; [3]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; [4]Alvaro-Fuentes, J., Cantero-Martinez, C., Lopez, M. V., Paustian, K., Denef, K., Stewart, C. E., & Arrœe, J. L. (2009). Soil Aggregation and Soil Organic Carbon Stabilization: Effects of Management in Semiarid Mediterranean Agroecosystems. *Soil Science Society of America Journal*, 73(5), 1519. doi:10.2136/sssaj2008.0333