Data Narratives: Increasing Scholarly Value

Line Pouchard
Purdue University
West Lafayette, IN, USA
pouchard@purdue.edu

Amy Barton
Purdue University
West Lafayette, IN, USA
hatfiea@purdue.edu

Lisa Zilinski
Purdue University
West Lafayette, IN, USA
lzilins@purdue.edu

ABSTRACT

Data narratives or data stories have emerged as a new form of the scholarly communication focused on data. In this paper, we explore the potential value of data narratives and the requirements for data stories to enhance scholarly communication. We examine three types of data stories that form a continuum from the less to the more structured: the DataONE data stories, the Data Curation Profiles, and the Data Descriptors from the journal *Scientific Data*. We take the position that these data stories will increase the value of scholarly communication if they are linked to the datasets and to the publications that describe results, and have instructional value.

Keywords

Datasets, data stories, research data management, scholarly communication, data narratives

INTRODUCTION

Data narratives or data stories have emerged as a new form of the scholarly communication focused on data. DataONE, the federated framework of Earth Sciences data centers, is publishing data stories in blogs and as part of its training and education modules (Rebich Hespanha, Menz, & Bragg, 2013). In the field of librarianship and information science, one finds the Data Curation Profiles that were developed to help librarians understand the needs and requirements of researchers for data curation and preservation (Witt, Carlson, Brandt, & Cragin, 2009). Other disciplines are also turning to narratives as an additional element useful to elucidate questions. In medicine, narrative inquiry focuses upon a single case, as opposed to evidence-based medicine that focuses on populations (Bleakley, 2005). In statistics, telling data stories that help interpret plots and make sense of the results has become an important goal of statistical education (Pfannkuch, 2010). In the publishing world, the publishing group, Nature, recently launched Scientific Data, an open access, peer- reviewed journal for the description of scientifically valuable datasets (http://www.nature.com/sdata/).

77th ASIS&T Annual Meeting, October 31- November 4, 2014, Seattle, WA, USA.

In this paper, we explore the potential value of data narratives and the requirements for data stories to enhance scholarly communication. We take the position that these data stories will increase the value of the research if they are structured and adhere to stringent criteria, or are peer-reviewed. If they are not, they have instructional value as they provide a human dimension to the data and research process.

The first part of this paper focuses on identifying different kinds of data stories. The second part of the paper examines the role of data stories and their potential impact. The third part discusses our findings and proposes some requirements that would increase their potential value. We leave data stories in medicine and statistics out of the scope of this paper, as they are discipline specific and serve a narrowly focused purpose.

DATA STORIES/NARRATIVES/DESCRIPTIONS

In this section we describe data stories that have different degrees of structuring, from the DataONE data stories, to the Data Curation Profiles, to the Scientific Data publication.

DataONE's Data Stories

DataONE data stories are the products of 30 to 60 minute semi-structured, audio-recorded interviews with 24 researchers that exemplify both satisfying and challenging experiences related to data (Rebich Hespanha, et al 2013). They vary widely in content, length, issues addressed, and purpose. Some stories go to great length to describe the data collection, storage and access mechanisms, intention to share, and some data pre-processing steps that help contribute to the quality assurance performed on the data. An example of such a story, "Tallying every bug and byte" recounts how a dataset counting insect species at over 40 sites over several years collected by a retiring professor piqued the interest of a graduate student who embarks on a project to curate, preserve, share the dataset and eventually deposit it in a national repository (Menz, 2014). This story introduces several characters: the graduate student Nora, her colleague Didi, and the retiring professor Andy. A basic overview about the dataset is provided, as well as Nora's transformative experience. It is an inspiring story that provides insight into the Data Scientist mental processes while working with the data. This story manages to convey both the practical experience of working with the data and the value added by data curation: without the work of Nora and Didi, several years of painstaking observations about a natural phenomenon would be lost, as the data owner, Andy, is retiring.

Another DataONE story, "Metadata? I thought you were in charge of that" describes how a team of ecologists who made the choice to re-use existing data discovered that metadata for many of their datasets was missing (RebichHespanha, 2013). This discovery was made at the deadline of a conference presentation and resulted in datasets (and their analysis) being excluded from the presentation. This occurred in spite of the careful planning in the collection and management of the datasets for re-use.

The first story mentions the existence of tangible results, a new dataset and a possible grant application. They both lack references to sources, publications, and to the dataset. None of the DataONE stories attracted comments from blog readers.

We now turn our attention to a different type of data stories, the Data Curation Profile (DCP).

Data Curation Profiles as Narratives

Unlike the DataONE stories, DCPs are the results of highly structured interviews of individual researchers about their data management practices in their research or about one particular project (Witt, et al, 2009). DCPs provide information about a particular dataset or collection of datasets, the steps the researcher is taking to manage or curate datasets, and what the researcher would like to do about the data, highlighting unmet needs. The DCPs are the results of collaborative work between data librarians, library liaisons, and researchers, including faculty or graduate students. The DCP Toolkit is openly available for download at http://datacurationprofiles.org/.

The interviews last between 60 and 90 minutes during which the researcher is also asked to fill out a detailed questionnaire about their data practices. Researchers' answers serve as entry points to a deeper conversation about specific data sets and data practices, data formats, storage, metadata, sharing and preservation. The interviews are transcribed and coded, and the obtained material is reported into the DCP, following a specific format prescribed by DCP curators. The resulting DCPs provide an in-depth description of the research process highlighting the data.

DCPs are intended as "stories" from inception: the welcome page of the DCPs Toolkit website describes a DCP as "essentially an outline of the story of a dataset or collection, describing its origin and life cycle within a research project" (*Data Curation Profiles Toolkit*, n.d.). It is described from the point of view of the researcher, and is meant to incorporate and convey the voice of the researcher. DCPs are submitted to publication and published with a DOI in the DCP directory but are not peer-reviewed. Instead, six criteria are given for acceptance of a DCP. It must be complete, i.e. describe a dataset and its lifecycle, provide information about current practice, and identify issues or areas of need.

Other criteria include the DCP's usability for other information professionals or librarians, the amount of information in the opening section to contextualize it, and a title describing the discipline/sub- discipline following the Classification of Instructional Programs nomenclature from the U.S. Department of Education's National Center of Educational Statistics.

A DCP may have several authors. These author(s) are the information professionals who conducted the interviews and wrote the profiles. The interviewee's name is withheld from the DCP publication. The DCP template, V1, also mentions data authors and data clients. Data authors generate the data, while data clients are the individuals who were interviewed and whose point of view is represented in the DCP. Data authors and data clients can be the same individuals if they generated and used their own data. Data clients are researchers who re-use data from existing repositories or external sources (*Data Curation Profiles Toolkit*, n.d.).

DCPs have a prescribed format found in the DCP Template for relating the information discovered in the interviews (*Data Curation Profiles Toolkit*, n.d.). The thirteen prescribed sections are described in Witt et al (2009) and include an overview of the research, data forms and stages, the value of the data, the data for ingest, and other descriptions. The template instructs the author(s) to leave no section blank, but note if no information is available on a particular topic.

We will now examine another type of data story: the articles published in "Data Journals" using the example of the *Scientific Data* (SD) journal.

Scientific Data's Data Descriptors as Scientific Products Data Journals are "a new format of publication that focuses on the dataset rather than the results of an analysis or the investigation of a scientific hypothesis". *Scientific Data* is an open-access, peer-reviewed publication for descriptions of scientific datasets initially focused on the life, biomedical and environmental science and the description of the experimental settings used to collect the data. Tests of new scientific hypotheses and extensive analysis are out of scope. The data described in the Data Descriptors must be deposited in community-recognized or public repositories (such as Figshare and Dryad). Datasets must be accessible to referees at the time of review and to the public at the time of publication. The main sections, metadata, include title, abstract, background and summary, methods, technical validation, usage notes, figures and tables.

VALUE OF DATA NARRATIVES

The DataONE Data Stories project was created to help "bridge the gap between current and potential data application" (Rebich Hespanha, et al, 2013). In a tweet by Hohmann, Carol Tenopir, in a presentation at the International Association of Scientific and Technological University Libraries' 2014 Annual Conference, said data and stories can increase the library's value (2014), and received a response by David Scherer "Agreed. We need more narrative driven messages that can also compliment data" (2014).

Linking data stories and narratives to datasets and journal articles in scholarly packages can increase and enhance the value and impact of the research. The data story or narrative can be part of supplemental information necessary for describing and providing context for the datasets that are analyzed in papers. Data stories and narratives:

- Increase the understanding of the data, the context of the data collection, and challenges encountered by the research team(s).
- Increase the probability of reuse and provides greater exposure to data.
- Increase the potential impact for national recognition and tenure/promotion through additional citations.
- Have instructional value for learning sound data management practices.

Evidence of value

Data stories and narratives take many forms such as those previously described. For example, they can also be included in supplemental materials as "expanded explanations of methods" (Kenyon Sprague, 2014) or "narrated short movies," (Howayek, et al, 2012). There is evidence that linking datasets to publications increases downloads suggesting greater impact for scholars (Zilinski, et al, 2014). Further, linking data stories as supplemental materials can play a role in data sharing (Kenyon Sprague, 2014). This suggests that when linked to the publication and dataset, these materials have the potential to increase data sharing, reuse, and validation.

DISCUSSION

Data Stories and narratives can provide context and additional description for a dataset that other supplemental material may not address.

Both DataONE stories are written in the past tense and were written from the point of view of an omniscient narrator. The anecdotal tone makes it easy to read and emphasizes the salient points. The introduction of the data scientist's persona and that of the other characters make it sound slightly artificial and the reader is left wondering if part of the story is fiction. However, the storytelling tone adds a human dimension to the data stories that makes them ring true to an audience: this enhances their instructional value.

The DataONE stories have value for teaching and engagement with data. As noted in the "Tallying every bug and byte" story, a graduate student was able to retrieve, curate and potentially re-use data. The story has the potential to instruct graduate students the importance of data curation. The "Metadata? I thought you were in charge of that" has the potential to instruct graduate students the importance of documenting data, and especially utilizing metadata. Forthcoming, DataONE will publish data stories with discussion questions within the narrative to encourage thoughtful conversations about data management in the context of the story (S. Rebich Hespanha, personal communication, May 23, 2014). The actual value of these stories is found in their learning potential rather than for inclusion in the scholarly package per se.

These two stories illustrate different aspects of the Data Life Cycle. The first one is clearly situated closer to the publication, as it is clear that the researcher has completed the project and is moving toward expanding it. The second one is situated at the beginning of the Data Life Cycle with the researcher documenting data collection. From the point of view of enhancing scholarly communication, these two stories provide insight into the research process but as they are not linked to datasets and they are not peer-reviewed, the scientific rigor required in scientific communication is lacking.

The Data Curation Profiles are highly structured and adhere to strict criteria for publication. The data stories gleaned from these publications give insight into researchers' interactions with and management of data throughout the data lifecycle. While perhaps not discussing a particular dataset, although it could, these data stories have instructional potential as well. The audience in this case is broader than just students. Beyond students, the DCP's can enhance data curators', librarians', and IT staff's understanding of gaps and needs that need to be addressed in the data lifecycle.

DCPs and DataONE stories are both the products of interviews with researchers. The resulting outcome differs due to the prescribed nature of the narrative for the DCPs. The structured, factual approach underlines the scientific method applied to obtain DCPs.

Data Descriptors are not the product of interviews. They are validated by the peer-review process. Thus they acquire credibility by the traditional scholarly communication method. In addition, the links to datasets deposited in publicly-accessible repositories enhance the validity and add value for potential re-use of the datasets. These publications can represent a significant step in the changing landscape of scholarly communications.

Data Descriptors are authored by the scientists involved in setting up the experiments and collecting the data. The researcher's voice is hidden under the factual description of the data and its context, in the traditional manner of scientific publishing. Their direct applicability for instructional purposes is limited because the human dimension of the data story is removed by the factual narrative.

CONCLUSION

In this paper, we have presented three types of data stories, the DataONE stories, the Data Curation Profiles (DCP), and the Data Descriptors from *Scientific Data*. Data stories are on a spectrum from having low credibility, but inclusive of a human dimension (DataONE), to credible, with structure and criteria (DCP), to validated, with peer-review and dataset deposit (Data Descriptors). We believe that data stories, across the spectrum, are of value. There is scholarly value in terms of providing, in a structured narrative, context and applicable information for the reuse of data. Stories with structure adhering to specified criteria and peer-reviewed stories fit into this category. Stories without these characteristics, such as the DataONE stories, add instructional value and can be used as a learning tool in data management.

REFERENCES

Bleakley, A. (2005) Stories as Data, Data as Stories: Making Sense of Narrative Inquiry in Clinical Education. *Medical Education, 4(3),* 543-40.

Data Curation Profiles Toolkit (n.d.). Retrieved June 10, 2014, from http://datacurationprofiles.org/

El Howayek, A., A. Bobet, S. Dawood, A. Ferdon, M. Santagata, and N. Z. Siddiki. "Project Implementation: Classification of Organic Soils and Classification of Marls—Training of INDOT Personnel'" Publication FHWA/IN/JTRP-2012/22. Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana, 2012.

Hohmann, T. [guacamole37]. (2014, Jun 05). #iatul2014 Tenopir. In order to convince re. the library's value you need dataand stories [Tweet]. Retrieved from https://twitter.com/guacamole37/status/4745194161606 4 9216

Kenyon, J., & Sprague, N. R., (2014). Trends in the Use of Supplementary Materials in Environmental Science *Issues in Science and Technology Librarianship, Winter (75),* doi: 10.5062/F40Z717Z

Menz, S. (2014). Tallying Every Bug and Byte. Retrieved from https://notebooks.dataone.org/data-stories/tallying-every-bug-and-byte/

Pfannkuch, M., Regan, M. Wild, C., Horton, N. (2010). Telling Data Stories: Essential Dialogues for Comparative Reasoning. *Journal of Statistics Education, 18(1),* 1-38.

Rebich Hespanha, S. (2013). Metadata? I Thought You Were In Charge of That. Retrieved from https://notebooks.dataone.org/data-stories/metadata-i-thought-you-were-in-charge-of-that/

Rebich Hespanha, S., Menz, S., & Bragg, J. (2013). *In Their Own Words: Researchers' Stories of Challenges and Triumphs in Data Management and Sharing*. Poster presented at the fall meeting of the American Geophysical Union, San Francisco, CA. doi: 10.6084/m9.figshare.892403

Scherer, D. [davidascherer]. (2014, Jun 05). @guacamole37 Agreed. We need more narrative driven messages that can also compliment data. #iatul2014 [Tweet]. Retrieved from https://twitter.com/davidascherer/status/474521805097 79 5584

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

Zilinski, L., Scherer, D., Bullock, D., Horton, D., & Matthews, C. (forthcoming). Evolution of Data Creation, Management, Publication, and Curation in the Research Process. *Transportation Research Record: Journal of the Transportation Research Board.*