DataUp: Enabling Data Stewardship for Researchers

Carly Strasser
University of California Curation Center, California Digital Library
carly.strasser@ucop.edu

Abstract

The move towards digital data is ubiquitous across all domains in academic research (Interagency Working Group on Digital Data, 2009; Carlson, 2006; C. L. Borgman, 2009; Faniel & Zimmerman, 2011; C. L. Borgman, 2008), and these data can be made available and distributed more quickly than ever before. This is often called the data deluge, and is a phenomenon that has been explored in the traditional academic literature (Carlson, 2006; Faniel & Zimmerman, 2011; C. Borgman, Wallis, & Enyedy, 2007), as well as in several major media outlets (Editors, 2010; Pollack, 2011; G Bell, 2009).

Among the most pressing problems associated with the data deluge is good data management: how does one handle the huge volume of available information effectively and efficiently to solve important problems? Knowledge of good data management techniques and software development lags behind the progression of the data deluge. Consequently, although researchers of all fields are faced with huge volumes of data from increasingly diverse sources, they do not have the skills to handle their data sets. This challenge is amplified by the fact that research data are seldom shared, re-used, or preserved (Nelson, 2009; Tenopir et al., 2011; LeClere, 2010). There is a growing awareness among practitioners and funders that this situation represents inefficient use of research dollars, missed opportunities to exploit prior investment, and a general loss for the scholarly community (Editors, 2009). Michener, Brunt, Helly, Kirchner, and Stafford (1997) described the loss of valuable data and insight about those datasets as “information entropy”. This loss of information is becoming increasingly worrisome as data management practices improve very slowly, while the volume of data grows exponentially.

Recognizing that most Earth, environmental, and ecological scientists use spreadsheets at some point in the life cycles of their data, the California Digital Library partnered with Microsoft Research Connec- tions and the Gordon and Betty Moore Foundation to create a tool for Microsoft Excel that would encourage and enable good data stewardship practices.

To optimize the tool, we first identified the needs of the community via surveys of researchers. We found that, on average, researchers had poor data management practices, were not aware of data centers or metadata standards, and did not understand the benefits of data management or sharing. We used the survey results to compose a list of desirable components and solicited feedback from the community to prioritize potential features. The result of this effort was a document outlining the requirements for DataUp; these requirements were made publicly available for comment on the DataUp blog, and were provided to developers who then created the DataUp software.

The resulting DataUp tool (dataup.cdlib.org) facilitates documenting, managing, and archiving tabular scientific data. It comes in two forms, both open-source: an add-in for Excel and a web-based application. The add-in operates within the well-known program, Excel; the web application allows users to upload tabular data to the web-based tool in either Excel format or comma-separated value (CSV) format. Both the add-in and the web application provide users with the ability to (1) Perform a “best practices check” to ensure data are well formatted and organized; (2) Create standardized metadata, or a description of the data, using a wizard-style template; (3) Retrieve a unique identifier for their dataset from their data repository, and (4) Post their datasets and associated metadata to the repository.

Acknowledgments: Microsoft Research Connections and the Gordon and Betty Moore Foundation provided funding for this work.
Strasser, C. (2013). DataUp: Enabling data stewardship for researchers. IConference 2013 Proceedings (pp. 657-658).
doi:10.9776/13300

Although there are hundreds of data repositories available to researchers for data archiving, the majority of scientists are not aware of their existence or how to access them. One of the major outcomes of the DataUp project is the ONEShare repository, created specifically for DataUp. Users can deposit their tabular data and metadata directly into the ONEShare repository from within the tool, allowing for seamless data archiving within the researcherÕs current workflow. An added advantage of ONEShare is its connection to the DataONE network of repositories (Michener, 2009). DataONE links together existing data centers and enables its users to search for data across all participating repositories using a single search interface. Data deposited into ONEShare will be indexed and made discoverable by any DataONE user, facilitating collaboration and enabling data re-use.

CDL envisions the future of DataUp to be directed by the participating community at large. Code for both the add-in and web application is open source[1] and participation in its improvement is strongly encouraged. Interested developers can expand upon and increase the tool's functionality to meet the needs of a broad array of researchers. To facilitate this effort, a list of requested improvements, bugs, and feature requests is maintained by CDL[2]. Although the target audience for the tools that result from the DataUp project will be Earth, environmental, oceanographic, and ecological scientists, we believe that any tools developed will be adaptable to other research communities, such as the social sciences.

Keywords: data management, research data, Excel, services, software tools

References

Borgman, C., Wallis, J., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries, 7*, 17-30.

Borgman, C. L. (2008). Data, disciplines, and scholarly publishing. *Learned Publishing, 21*, 29-38.

Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities, 3*(4).

Carlson, S. (2006). Lost in a sea of science data. *The Chronicle of Higher Education, 52*(42), A35.

Editors. (2009). Data's shameful neglect. *Nature, 461*(7261), 145.

Editors. (2010). The data deluge: Business, governments and society are only starting to tap its vast potential. *The Economist*, 25 Feb.

Faniel, I., & Zimmerman, A. (2011). Beyond the data deluge: A research agenda for large-scale data sharing and reuse. *The International Journal of Digital Curation, 6*(1), 58-69.

G Bell, A. S., T Hey. (2009). Beyond the data deluge. *Science, 323*(5919), 1297-1298.

Interagency Working Group on Digital Data. (2009, January). *Harnessing the power of digital data for science and society* (Tech. Rep.). Committee on Science of the National Science and Technology Council.

LeClere, F. (2010, August). *Too many researchers are reluctant to share their data*.

Michener, W. (2009). DataNetONE: Observation Network for Earth. (NSF Grant No. OCI 0830944).

Michener, W., Brunt, J., Helly, J., Kirchner, T., & Stafford, S. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications, 7*(1), 330-342.

Nelson, B. (2009). Data sharing: Empty archives. *Nature, 461*, 160-163.

Pollack, A. (2011). DNA Sequencing Caught in the Deluge of Data. *New York Times, Published 30 November*.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*(6), e21101+.

[1] http://bitbucket.org/dataup/main

[2] http://bitbucket.org/dataup/main/wiki/improvements_issues