# Datastar – A Semantic Registry for Research Datasets

Dean B. Krafft, Kathy Chiang, Jon Corson-Rikert, Huda Khan, Wendy Kozlowski, Brian Lowe, Leslie McIntosh, Mary Ochs, Gail Steinhart, and Sarah Wright

Cornell University Library

With the growing mandates for open access to the outputs of research, including research data[[1]](#footnote-1), there is a critical need for systems that support locating, accessing, and using research datasets – ideally in the context of their associated publications, grants, and researchers. Datastar is an open-source, semantic-web-based platform supporting the description, discovery, access, curation, and reuse of research datasets. Datastar extends the ability of the VIVO researcher profiling system[[2]](#footnote-2) to represent relationships among researchers, grants, and publications to capture the scholarly context around research datasets and to highlight dataset citations. Datastar can be run either as a standalone dataset registry or as an extension to an institutional instance of VIVO.

The Datastar project was initially funded in 2007 by NSF[[3]](#footnote-3), and more recently in 2011 by the U.S. Institute of Museum and Library Services. In this presentation, we will report on the more recent work, which has involved both working with researchers to create a set of Data Curation Profiles to understand researchers’ needs and preferences with respect to documentation, sharing, dissemination, and reuse of datasets; and applying those profiles to the development of the Datastar application. The current version of the open source Datastar software is available on SourceForge at <http://sourceforge.net/p/vivo/datastar/>.

The proposed presentation will cover the following specific topics:

* Our use of the Data Curation Toolkit to conduct interviews and create a set of Data Curation Profiles (DCPs) with researchers at Cornell University and Washington University, St. Louis.
* How the identified researcher priorities in areas such as internet discoverability, creation of a basic, public description, and data citation guided the development of the Datastar application.
* The development of a Datastar ontology that describes datasets with sufficient context to facilitate review by researchers and to enable the discovery of datasets across domains.
* How Datastar extends the VIVO architecture and integrates with VIVO to represent research datasets and explicitly relate them to researchers, grants, publications, and other datasets.
* A brief overview of the Datastar interface.
* How Datastar makes extensive information about research datasets available as publicly shareable Linked Open Data for discovery, harvest, and reuse by other systems.
* How Datastar fits into the overall workflow of research dataset description, deposit in repositories such as Cornell’s DSpace-based eCommons or Fedora-based CULAR, curation, preservation, citation, and reuse.

The Datastar project and open source application software address several aspects of the OR2013 theme of Use, Reuse, and Reproduce. Datastar promotes open access and easy discoverability of research datasets; it promotes reproducible research through interlinking research datasets with the full research context of researchers, publications, and grants; and it complements the capabilities of open institutional and disciplinary repositories as part of the overall research ecosystem.

1. <http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research> [↑](#footnote-ref-1)
2. <http://vivoweb.org> [↑](#footnote-ref-2)
3. <http://www.nsf.gov/awardsearch/showAward?AWD_ID=0712989&HistoricalAwards=false> [↑](#footnote-ref-3)