DataStaR: A Data Sharing and Publication Infrastructure to Support Research

Gail STEINHART

Albert R. Mann Library, Cornell University, USA, GSS1@cornell.edu

Abstract

DataStaR, a Data Staging Repository (http://datastar.mannlib.cornell.edu/) in development at Cornell University's Albert R. Mann Library, is intended to support collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centers and/or institutional repositories. Researchers may store and share data with selected colleagues, select a repository for data publication, create high quality metadata in the formats required by external repositories and Cornell's institutional repository, and obtain help from data librarians with any of these tasks. To facilitate cross-domain interoperability and flexibility in metadata management, we employ semantic web technologies as part of DataStaR's metadata infrastructure. In this paper, we describe the overall design of the system, our work to date with Cornell researchers and their data sets, and possibilities for extending DataStaR for use in international agriculture research.

Introduction

Sharing research data facilitates collaborative research among colleagues, and, when shared more widely, holds the potential to advance progress within a given discipline and even across disciplines. Research data may be used to reproduce and verify past results, plan future experiments, and support comparative studies and meta-analyses.

In spite of the potential benefits of sharing data, barriers to sharing exist. These may be cultural or sociological (sharing may not be the norm in certain disciplines; individuals may fear being "scooped"), procedural (confidentiality or commercialization concerns may mitigate against sharing), technological (suitable and accessible infrastructure may simply not exist), or logistical (researchers lack the skills and/or time to share their data). Our intention with the DataStaR - short for "Data Staging Repository" – project (http://datastar.mannlib.cornell.edu/) is to support the research process in a way that encourages data sharing more widely, primarily by reducing the significance of the last two barriers.

DataStaR: what is a Data Staging Repository?

DataStaR is both a platform and a set of services meant to facilitate data sharing in a way that is controlled by the researcher, as well as publication of data and metadata to appropriate repositories. We focus primarily on support for so-called "small science" data sets, those that don't require specialized infrastructure for storage, management, and access. DataStaR itself is only a temporary repository for data – working versions to be shared by colleagues, or final versions in preparation for submission to a permanent data repository. The notion of an intermediate working repository has other precedents. An infrastructure to support a curation continuum for research data, consisting of private, collaboration, and publication domains has been described and developed at Monash University in Australia by Treloar et al. (2007). Green and Gutmann (2007), in a paper describing the possibilities for partnerships between institutional repositories and domain-specific repositories to encourage the migration of data

from local to more widely shared environments, also get at this notion of a continuum or progression, and the types of support required to move works in progress to published versions. We emphasize that DataStaR is not a preservation repository, but is managed with long-term preservation of research data in mind (Steinhart *et al.* 2009).

There are multiple benefits to the staging repository model. For users, DataStaR offers a managed and controlled environment for collaboration with selected colleagues, off-site back up of valuable research data, tools to create metadata in a variety of formats, the ability to reuse information from previously created metadata, and assistance from librarians in determining an appropriate publication strategy and preparing data and metadata for publication. For librarians concerned with promoting responsible custodianship of research data created at their institution, the arrival of a new data set in DataStaR signals a curatorial opportunity. We see this combination as a potentially successful way to support the research process while simultaneously encouraging and supporting the publication of data sets to permanent repositories.

The DataStaR system consists of a Fedora-based repository (http://fedoracommons.org/) for storage of data set files, a semantic metadata store based on the vitro software (http://vitro.mannlib.cornell.edu/) - a web-based ontology and instance editor developed at Mann Library, additional open-source components (DROID for file format identification, http://sourceforge.net/projects/droid/ and SWORD for deposit to some repositories, http://www.swordapp.org/), as well as custom code written specifically for this project (Figure 2). A user may interact with the DataStaR system in the following ways: a researcher may upload a data set to the DataStaR repository, create minimal metadata, and assign permissions to grant access to data and metadata to selected colleagues, or the general public. At the time of upload, the user must indicate a destination repository for publication, although "to be determined" is a valid selection in the event a user is undecided or intends to use DataStaR solely for sharing data and not for publication. If no repository is selected, the user is presented with a simple (and optional) form for additional metadata. The selection of a specific destination repository triggers the display of a metadata form appropriate to that repository, although completion of this form is not required until the user is ready to publish the data set. Prior to or at the time of publication, the user completes the required metadata, consulting with project librarians as needed. The specifics of how a data set moves from DataStaR to a destination repository are varied, depending on the submission mechanism for that external repository. In some cases we are able to support direct deposit from DataStaR; in others, human mediation is necessary.

To facilitate cross-domain interoperability and flexibility in metadata management, we employ semantic web technologies as part of DataStaR's metadata infrastructure. Briefly, existing metadata schemas are converted to OWL ontologies and incorporated into the DataStaR system. An advantage for users is that treating metadata as a collection of statements rather than static and stand-alone documents facilitates the reuse of previously created statements in new metadata. It's not at all uncommon for a researcher to use the same field or laboratory methods, for example, or to conduct multiple studies in the same geographic location. Once that information has been entered in DataStaR to describe one data set, it is easily reused in the description of others. More broadly, we aim to support linked data in the future, and operate on the assumption that increasingly the application of semantic web approaches and technologies to the management of metadata will become standard practice. Our motivation for and approach to implementing a semantic approach to metadata management is described more fully in Lowe (2009).

In the first phase of development, currently underway, it's our goal to support publication to the repositories listed in Table 1. These repositories and standards reflect the needs of the researchers we work with; support for additional repositories and metadata standards may be added later, and will reflect the demands of the researchers with whom we work.

Providing data curation services to Cornell researchers

Currently, the DataStaR team is working with a number of research groups and individual researchers. These include Cornell's Upper Susquehanna River Basin Agricultural Ecology Program, the Cornell Biological Field Station, the Cornell Plantations Natural Areas Program, the Cayuga Lake Watershed Network, the Loon Project, and the Virtual Center for Language Acquisition. In addition, we plan to use DataStaR as a submission mechanism for data sets contributed to the Cornell University Geospatial Information Repository (CUGIR, http://cugir.mannlib.cornell.edu/).

The researchers involved with DataStaR are already motivated to share their data. Their motivations vary; in the case of the Upper Susquehanna River Basin Agricultural Ecology Program and the Cayuga Lake Watershed Network, sharing is motivated in part by a desire to make scientific findings available to managers, policy makers, and the general public. The Loon Project, as a recipient of funding from the US National Science Foundation's Long Term Research in Environmental Biology program, is explicitly required to disseminate its data, and is making use of DataStaR to do so. Long-term research also motivates sharing for the Cornell Biological Field Station, the Cayuga Lake Watershed Network, and the Cornell Plantations Natural Areas Program. Here, sharing previously collected data enables new research – whether by facilitating analysis over time, or simply providing background information to guide new research efforts. Facilitation of collaborative work within a research group is also a motivator: well-documented data, centrally accessible, makes it easier for collaborators to reuse and integrate data collected by others into their own research. In the case of the Virtual Center for Language Acquisition, this allows collaborative analysis of audio recordings, while in the case of the Upper Susquehanna River Basin Agricultural Ecology Program, sharing allows simulation modelers to validate their models using field-collected data. Because we tend to work with groups already predisposed to share data, for a variety of reasons, we don't tend to encounter resistance to sharing.

While our collaborating researchers are motivated to share, they do have questions or concerns about the process, and most appreciate some level of assistance. The most commonly needed forms of assistance include help in deciding which data to share, help with data organization and formatting, and help with metadata creation. Deciding which data to share and how it should be organized depend to some extent on anticipated uses. For researchers collecting environmental data, anticipated reuse usually means analyzing data over time, or combining data sets from multiple researchers to perform comparative analyses. In these cases, data sets where the data have been somewhat processed are usually the most useful, rather than the raw data themselves, although raw data may allow others to check a researcher's intermediate calculations and final results. Decisions about organizing data usually involve trade-offs that affect ease of use for the end user and ease of preparing and updating the data for the data owner. File format decisions also sometimes involve tradeoffs. Current and common proprietary formats may be easy to create and use, but are not suitable for long-term preservation, and may be incompatible with software other than that with which they were created. Non-proprietary formats, such as tab- or comma-delimited text files for tabular data, while they may not be in the working format that a researcher is accustomed to, are more stable in terms of long-term preservation and have greater potential for cross-platform compatibility now and in the future.

In terms of metadata, some metadata elements are fairly easy to understand and complete. Others may require specialized knowledge or an eye for details that researchers might reasonably overlook. Some examples of areas where we've provided expertise to researchers in completing metadata include the use of controlled vocabularies for keywords and subject terms, assistance with crafting language for intellectual rights statements, and adherence to established conventions for specifying geographic coordinates.

Conclusion and prospects for applications in international agriculture

While we have worked with only a handful of research groups, we're pleased with their response to our services, which seem to fill some very real needs at Cornell. Since the project began, several researchers have asked us for assistance with data archival and dissemination plans in grant proposals. Furthermore, the University of Melbourne is in the process of adapting the DataStaR software for use as a data registry for the Australian National Data Service, having already implemented the core vitro software as an expertise directory at the University of Melbourne.

We'd like to consider whether any elements of DataStaR, conceptual or technological, would be useful to the international agriculture research community. We recognize the existence of well-established systems such as the International Crop Information System (ICIS, Fox and Skovmand 1996), and its crop-specific instances such as the rice (IRIS) and wheat (IWIS) systems (e.g. McLaren *et al.* 2005). Systems like ICIS have certain advantages for agronomic data, including some degree of standardization of data that facilitates interoperability among data sets, and tools and applications for data use and analysis - capabilities that DataStaR lacks because it was developed to manage much more heterogeneous data sets. Nevertheless, we're interested in exploring whether any aspects of DataStaR might be useful in some research contexts: infrastructure for preliminary and controlled data sharing during the research process, tools for documenting and moving data from that preliminary (staging) environment into the publication domain, and semantic approaches to metadata management. Understanding its applicability requires consideration of one or more questions in each of these areas.

While we're not deeply familiar with the norms and practices for research in this area, we speculate that a data staging repository might work well to facilitate collaboration among researchers, particularly if they are distributed geographically, and if internet connectivity is somewhat reliable. A shared data repository can make it easier for researchers to ensure they are working from the same version of a data set, and also serve as a remote back-up.

There are two main requirements for the documenting and publishing function of DataStaR to be useful in this arena. The first is the existence of suitable destination repositories for agricultural research data. For researchers already participating in efforts such as ICIS or other systems, there may be no particular advantage to adding an intermediate layer to the process of publishing data, above and beyond the opportunities afforded by sharing works in progress with selected colleagues in the pre-publication stage. We don't know whether there are other repositories, institutional or discipline-specific, that might be usefully linked to a staging repository. If none exist, a single repository could serve both the staging and publication functions. The second requirement has to do with assistance to researchers in preparing and documenting data sets for publication. The reason this intermediate infrastructure works well at Cornell is that staff are available to guide researchers through the process of documenting and preparing data for publication, and the staging repository works well as a shared workspace where this preparatory work can be accomplished. Support for researchers using the system is as important a part of the system as the technological infrastructure itself.

Finally, a semantic approach to metadata management is, technologically, one of the most innovative aspects of DataStaR. The immediate benefit to users is the ability to reuse their own information, rather than re-entering or copying statements from one metadata record to another. Additional benefits of managing metadata in this way are more likely to be realized in the future, when linked data and infrastructure to support it are more common. Early adopters, while perhaps incurring some additional overhead in supporting a somewhat more complex infrastructure (compared to implementing an existing out-of-the-box repository solution) will avoid future costs of retrospective conversion of metadata to support the semantic web. The immediate utility of this approach depends on whether a

flexible data sharing and publication environment capable of supporting multiple standards has value for the community, or whether useful links to other agricultural information systems already employing semantic web technologies can be made.

Acknowledgements

The author gratefully acknowledges the contributions of the DataStaR team: Brian Caruso, Kathy Chiang, Jon Corson-Rikert, Dianne Dietrich, Ann Green, Brian Lowe, and Janet McCue, as well as Mary Ochs for providing comments on a draft of this paper.

This material is based upon work supported by the National Science Foundation under Grant No. III-0712989. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

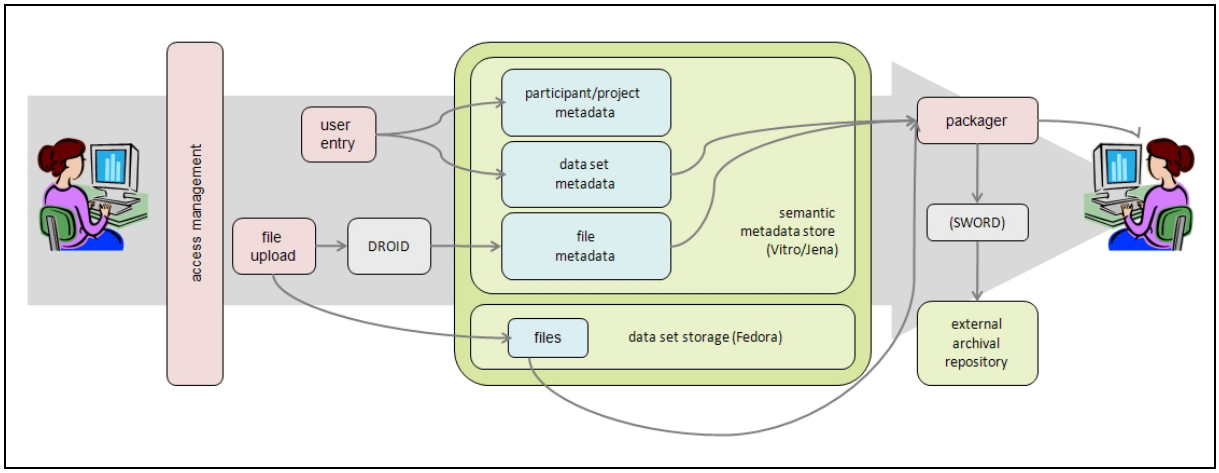

Fig. 1. Overview of DataStaR's system architecture, from left – initial access to the system by a data owner, to right – publication and distribution of data to other users. An access layer controls who may access the system and gives users the ability to grant access to others for their content. Users enter metadata about themselves and their research group as well as metadata for their data sets. The format of uploaded data files is determined by DROID and stored, along with other file-specific information, in the semantic metadata store, while data files are stored in a Fedora repository. For publication and direct distribution to users, XML metadata is written from the semantic metadata store. Data and metadata are downloaded or transmitted to users or archival repositories directly, or, in the case of some repositories, via the SWORD protocol.

Table 1. The repositories/domains and their metadata requirements supported in DataStaR's first round of development.

Repository or domain	**Metadata requirements**
eCommons (Cornell's institutional repository, http://ecommons.cornell.edu/)	DSpace/Dublin Core
Cornell University Geospatial Information Repository (CUGIR, http://cugir.mannlib.cornell.edu/)	Content Standard for Digital Geospatial Metadata (FGDC-CSDGM, http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata)

Knowledge Network for Biocomplexity (KNB, http://knb.ecoinformatics.org/)	Ecological Metadata Language (EML, http://knb.ecoinformatics.org/software/eml/)
Virtual Center for Language Acquisition (VCLA, http://vcla.clal.cornell.edu/)	Open Language Archives Community (OLAC, http://www.language-archives.org/)

References

FOX, P. N., and B. SKOVMAND. 1996. The International Crop Information System (ICIS) - Connects Genebank to Breeder to Farmer's Field. In: *Plant adaptation and crop improvement*. M. Hammer and Cooper G.L., eds. Wallingford, UK: CAB International. p.317-326.

GREEN, A.G., and M.P. GUTMANN. 2007. Building Partnerships Among Social Science Researchers, Institution-Based Repositories and Domain Specific Data Archives. *OCLC Systems & Services* 23(1):35-53.

LOWE, B. 2009. DataStaR: Bridging XML and OWL in Science Metadata Management. *Metadata and Semantic Research* 46: 141-150.

MCLAREN, C. G., R. M. BRUSKIEWICH, and A. M. PORTUGAL, A. B. COSICO. 2005. The International Rice Information System. A Platform for Meta-analysis of Rice Crop Data. *Plant Physiology* 139(2):637-42.

STEINHART, G., D. DIETRICH, and A. GREEN. 2009. Establishing trust in a chain of preservation: The TRAC Checklist Applied to a Data Staging Repository (DataStaR). *D-Lib Magazine* 15(9/10).

TRELOAR, A.. D. GROENEWEGEN, and C. HARBOE-REE. 2007. The Data Curation Continuum: Managing Data Objects in Institutional Repositories. *D-Lib Magazine* 13(9).