DataStaR: An Institutional Approach to Research Data Curation

*Gail Steinhart**

Introduction and background

DataStaR, a Data Staging Repository (http://datastar.mannlib.cornell.edu/) hosted by Cornell University's Albert R. Mann Library, was conceived of as a platform and a set of services to facilitate data sharing among collaborators, and to enable faculty to publish digital data and high quality metadata to domain-specific repositories and institutional repositories. DataStaR is intended to serve as a temporary repository for data sets (in any stage of completion) that researchers may share with selected colleagues, as well as a platform with tools for creating metadata in variety of formats, supported by librarian-curators prepared to assist researchers in preparing and submitting both data and metadata for publication to an external repository for the long term.

There is ample evidence that even when appropriate data repositories exist for a particular discipline, researchers often fail to take full advantage of them (Glover et al. 2006, Karasti et al. 2006, Lord and Macdonald 2003, Martinez-Uribe 2008). In some disciplines, no such repositories exist, and researchers have few or no options to archive or to share their data. This lack of participation in data sharing and archival activities suggests an opportunity for academic libraries to provide a much-needed service. While DataStaR is not specifically focused on social science data, we offer it as a model of possible interest to data curators or archivists in any discipline.

Mann Library has some well-established data distribution activities, which include the Cornell University Geospatial Information Repository (CUGIR), the USDA Economics, Statistics and Marketing Information System (USDA-ESMIS), both domain-specific data repositories, as well as a completed NSF-funded project examining the possibilities for the library to collaborate with researchers to document and archive data. The last project entailed a great deal of highly-customized work to handle a relatively small amount of data, and we became interested in exploring more sustainable approaches that would also be more portable across different research groups and disciplines. We were also interested in expanding the arena in which we work to include supporting the collaborative nature of research as it progresses in real time.

Indeed, there is significant interest in exploring the possibilities for interaction between librarians and researchers throughout the entire information life cycle, as well as the role institutional repositories can play in distributing research data. Green and Gutmann (2007) make a compelling case for cooperation between institutional repositories and domain repositories to encourage the movement of data from one to the other. Treloar et al. (2007) describe a curation continuum that acknowledges a distinct collaboration (pre-publication) and a more formal publication and preservation realm; the boundary between the two suggests a process for migrating material from one to the other. D. Scott Brandt (2007) of Purdue University advocates librarian involvement further "upstream" in the research process, aligning and developing services to support the work of data management, documentation, publication, and preservation as the original research itself occurs.

Researchers that we've collaborated with thus far have had questions, concerns, or needs that give us some confidence that upstream partnerships between librarians and researchers can yield substantial benefits. The need for a collaborative space where researchers may share data with selected colleagues is clear, and Mann Library has received multiple requests for this type of support. Researchers have also asked for guidance as to how to make data related to a published journal article available in cases where the journal itself has no mechanism for distributing digital data sets. Researchers who are prepared to share data often have questions as to which data they should make available – raw or processed? Complete, or summarized? Finally, several researchers have expressed the desire that users let them know how they intend to use the data. This is for a variety of reasons; some researchers simply want to know what others are using their data for. Others are concerned that complex data may be misinterpreted or put to a particular use for which the data are not well-suited. Still others would like subsequent users to acknowledge the original source of funding in any resulting publications, and to be able to report information on use of their data to funding agencies. While the DataStaR project may not be able to address all of these concerns, these examples illustrate some of the issues that make developing a local, staging approach to data publication a compelling idea as well as a challenge to implement.

Model for a Data Staging Repository

In a conventional repository model, data producers typically interact with a data repository when research is complete. At the close of a project, data sharing with collaborators has most likely already taken place; the primary functions of the repository are to curate and preserve data, and disseminate it to end users. In our staging repository model, researchers may deposit data earlier in the research process without necessarily exposing it to the public. This serves a variety of purposes. A managed workspace allows for controlled sharing with selected colleagues (and the public, if the researcher desires), and remote storage and backup of data. Data sets intended for publication may also be submitted to the staging repository, and fully documented data sets meeting the requirements of external "destination" repositories may be passed from the staging repository for publication. Published data sets are deaccessioned from DataStaR. Unless the data set owner chooses to remove it, a metadata record linking to the external repository remains to facilitate discovery.

To illustrate the utility DataStaR may have for a researcher, consider the following example (Figure 1). An ecologist examining the spread of an invasive species in New York State compiles field observations of occurrences of the species of interest over time. She enters her preliminary data into a spreadsheet that she wishes to share with colleagues at a natural history museum. She uploads the spreadsheet to DataStaR, which automatically generates minimal metadata based on her account information and information about the file(s) that can be determined automatically during the ingest process. The researcher optionally completes additional metadata, and assigns her colleagues permissions to view metadata and download data. Later in the research process, she consults with a data librarian about publishing her data to a data repository to be managed over the long-term. By now, her original data set has grown and she has also generated GIS data sets that show the distribution of the invasive species at selected points in time. She and the librarian agree to deaccession the original research data, document and submit the final version of the complete observational data to an ecological data repository, document and submit the GIS data to a state-level GIS data clearinghouse, and to deposit copies of both data sets (and their detailed metadata) in Cornell's institutional repository. The researcher uses tools in DataStaR to add to the minimal metadata records for these new data sets, according to the different standards of the ecological and GIS data repositories. The mechanics of publication to an external repository vary, and depend on the architecture and policies of that repository. When the researcher publishes her data sets from DataStaR to Cornell's institutional repository, DataStaR extracts the needed metadata from the domain-specific records she created earlier, populates a metadata record in the institutional repository, and deposits both the data and domain-specific records (as supporting documents). Publication to the ecological data repository from DataStaR is also seamless and accomplished without human intervention. Publication of the GIS data to the state clearinghouse, because its submission process requires human mediation, is handled by the data librarian.

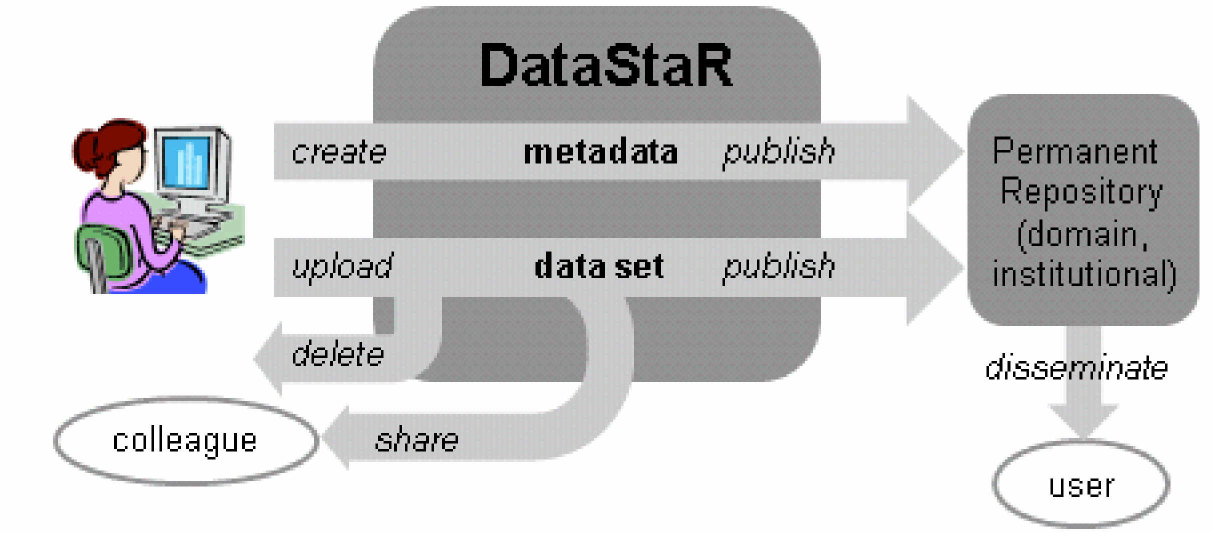


Figure 1. Interaction between a researcher and DataStaR.

Allowing the staging repository to be used as a collaborative workspace may complicate the flow of information. Researchers may share data that are never published to a permanent repository in their original form; instead, such data sets may be de-accessioned and replaced with publication-ready data sets. Consumers of data residing in the staging repository may be known collaborators or anonymous users, depending on the permissions specified by the data owner. Both the staging repository and domain or institutional repositories act as custodians of data in their possession and perform curation actions such as checksums, backups, etc. Adding DataStaR as another "layer" in the publication process may seem to complicate the process of publishing data to a permanent repository, and for some users this "layer" may be unnecessary, but we believe that the DataStaR repository and related support offer enough additional services and functionality to make the arrangement worthwhile and productive for many researchers

A novel approach to managing metadata

Another component of the DataStaR project is its planned metadata infrastructure. Information managers are increasingly interested in applying semantic web principles and technologies to metadata to support interoperability and machine processing (Bermudez and Piasecki 2006). We plan to "lift" existing metadata schemas into Web Ontology Language (OWL) ontologies, incorporating them into a growing assemblage of ontologies in the DataStaR system that will make it possible to treat metadata as a collection of discrete statements rather than as a standalone document. Users (data owners) may then reuse and recombine these statements to create new metadata without repetitive entry of information that is common to the description of multiple data sets. A consistent interface in combination with support for multiple metadata schemas will result in a system where users will be able to create metadata in different standard formats without having to be expert in each one. The system is also able to store and harness information about users and research groups in DataStaR in such a way that the work of a group can be displayed coherently on the DataStaR web site, or made available to that group for display on the group's own project web site.

Our earlier example of how a particular researcher might interact with DataStaR to publish data and metadata demonstrates how this approach to metadata might streamline dataset publication. The de facto standard for documenting GIS data is the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata , and this is the standard typically required for publication to a GIS data center in the United States. Ecological Metadata Language (EML) is the better choice for other types of ecological or environmental data; it is the standard required for publication to the Knowledge Network for Biocomplexity (KNB) . While EML can be used to document GIS data sets and GIS data may be deposited with the KNB, should a researcher wish to publish a GIS data set to a GIS data repository in the United States, they would likely be required to use the FGDC standard. It's unlikely that a researcher would be expert in both standards, but DataStaR would make it possible for a single researcher to create records easily in either format, reusing information as needed. In cases where a user may want to convert a complete record to another format, we plan to utilize existing crosswalks where they exist (EML to FGDC), and to create ontology-based crosswalks when they don't. This approach would also be used to facilitate deposit in Cornell's institutional repository (eCommons@ Cornell), which currently operates on the DSpace platform. The system extracts DSpace metadata from the collection of statements created for a domain-specific record and stored in DataStaR, thus minimizing the work involved in depositing to more than one repository. To avoid "losing" the rich information contained in a domain-specific metadata record, project managers encourage researchers to deposit that metadata record in eCommons as well, as supplemental material.

To implement semantic web technologies for the creation of metadata in DataStaR, the project team is extending Vitro , a Java web application that has supported several projects at Mann library. Vitro provides a customizable front end for searching and browsing a semantic graph of data, along with an interface for editing ontologies and instances. Mann Library has developed and deployed Vitro for an ontology-based web application bringing together the diverse research and education activities of faculty in the life sciences at Cornell (VIVO , Devare 2007).

DataStaR's role as a partner in digital preservation

DataStaR is intended to be a transitory home for research data sets, although one of the goals of the approach is to promote the movement of data to long-term preservation repositories. Toward that end, DataStaR administrators aim to be responsible partners in a process that leads to the preservation of research data sets by applying selected best practices in digital preservation to the staging repository context. The project team completed a process of evaluating and identifying criteria from the Trusted Repository Audit and Certification: Criteria and Checklist (TRAC) for their relevance to DataStaR, and compiled a set of documents to guide the creation of policies, system functional requirements, metadata requirements, and data management processes throughout the life cycle of DataStaR's digital resources

An institutional repository for research data?

DataStaR's primary constituents are Cornell researchers and their colleagues, as is the case for more traditional institutional repositories. However, with its emphasis on promoting the publication of data to permanent repositories, DataStaR also differs from institutional repositories and domain repositories in some important ways. Table 1 summarizes some selected characteristics of both types of repositories and DataStaR, highlighting differences and shared characteristics. DataStaR shares several key characteristics with the majority of institutional repositories: a focus on a local constituency, a relative lack of specialized tools and services for using or analyzing repository content (such as tools for visualization, extraction, and analysis of data) or other support for end users of repository content, and a lack of a specific deposit mandate. In other respects, DataStaR more closely resembles a domain-specific repository (albeit with the potential to serve multiple domains). Information managers originally engineered institutional repositories to handle text rather than data, while some domain repositories are focused on data. Domain repositories usually utilize a single, specific metadata standard for that domain, while DataStaR supports multiple domain-specific standards, as needed by its users. Domain repositories may have specific data formatting requirements, which DataStaR staff help researchers comply with in the process of publication of a data set. Finally, DataStaR staff provide significant support for data owners in preparing and submitting their data, as do some domain repositories. DataStaR is different from both institutional and domain repositories in that it is not a preservation repository, although as we noted earlier, a goal of the DataStaR project is to be a responsible partner in the preservation process

Current status and future work
Partnering research groups or individuals at the inception of the DataStaR project included the Cornell Language Acquisition Laboratory (CLAL) and the Upper Susquehanna River Basin Agricultural Ecology Program at Cornell University (USAEP , Woodbury et al. 2008). Our collaboration with the USAEP group has been quite active, with DataStaR facilitating the publication of several of the group's data sets to both the KNB and eCommons@ Cornell. Two new research group partners are the Cornell Biological Field Station (CBFS) and the Cayuga Lake Watershed Network. CBFS, located on the south shore of Oneida Lake, NY, serves as a primary field site for aquatic research at Cornell University. A 50-year long-term database on the food web of Oneida Lake that has been used in hundreds of publications is the centerpiece of the research program. Currently, DataStaR staff are collaborating with CBFS staff to prepare their long-term data for archiving and to create detailed, high-quality metadata. The Cayuga Lake Watershed Network is a community organization comprised of citizens, businesses, associations, and local governments throughout the Cayuga Lake Watershed. The Watershed Network promotes support for maintaining and improving the ecological health and beauty of the watershed, along with a healthy economy, in order to sustain a healthy social environment for watershed residents. The Watershed Network fulfills this mission by facilitating the discovery and exchange of information, including data collected by the many groups that are active within the watershed. In addition to these research groups, we have also been approached by individuals at Cornell with personal data collections each would like to archive. Finally, we aim to use DataStaR as a submission mechanism for GIS data being deposited to CUGIR, the GIS data repository maintained at Mann Library.

The DataStaR platform itself is still very much in development. We have identified and integrated the required metadata elements needed in order to manage data within the repository, and have integrated the EML ontology to support the creation of EML records. Even at this early stage, however, we've learned several important lessons. The first is that because we don't have a user-friendly system in place for non-expert users, the process requires a very high level of service (we must do most of the work for them) that data owners willingly accept. While we're reluctant to pass up opportunities to recruit new partners made possible by providing this level of service, we also recognize that doing so is not a sustainable approach. New challenges may arise as we try to encourage researchers to assume more of the responsibility for preparing data and metadata for submission to a repository. We've also learned even "low" barriers to a particular technology may not be as low as we'd like. For example, we've offered wikis to research groups to use, and while some groups use them effectively, others barely use them at all. In some cases faculty may have some other workaround with which they're already comfortable, and implementing a change in practice may be especially difficult in those situations. This observation will likely hold true as we encourage researchers to make use of the tools offered in DataStaR. In spite of these potential difficulties, we've found that in principal, researchers are quite ready to embrace the idea of making data publicly available, creating high quality documentation, and preparing data so that it remains usable well into the future. We've been invited to collaborate on grant proposals (contributing language on plans for archiving and distributing data), have collaborated on one journal article so far, and have been approached by various groups and individuals with an interest in sharing data.

Acknowledgements
The author gratefully acknowledges the work of the DataStaR development team: Brian Caruso, Kathy Chiang, Jon Corson-Rikert, Dianne Dietrich, Ann Green, Brian Lowe, and Janet McCue, 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.

*Contact: Gail Steinhart, Research Data & Environmental Sciences Librarian, Albert R. Mann Library, Cornell University, Ithaca, NY 14853, Phone:

607-255-7251.E-mail: GSS1@cornell.edu

References

Bermudez, L, & Piasecki, M. (2006). Metadata Community Profiles for the Semantic Web. Geoinformatica 10(2): 159-176. Online: http://www.springerlink.com/content/v72507q376750334/. Retrieved 9/30/2008.

Brandt, D. Scott. 2007. Data, research, metadata, metaresearch. Washington, DC, http://www.ala.org/ala/acrlbucket/stsconferencepro/annual2007programs/brandt.pdf. Retrieved 07/23/2008.

Devare, Medha. 2007. VIVO: Connecting people, creating a virtual life sciences community. D-Lib Magazine 13(7/8). Online: http://www.dlib.org/dlib/july07/devare/07devare.html. Retrieved 9/30/2008.

Glover, David M., Cynthia L. Chandler, Scott C. Doney, Ken O. Buesseler, George Heimerdinger, J. K. B. Bishop, and Glenn R. Flierl. 2006. The US JGOFS data management experience. Deep-Sea Research Part II: Topical Studies in Oceanography 53, (5-7): 793-802.

Green, Ann G., and Myron P. Gutmann. 2007. Building partnerships among social science researchers, institution-based repositories and domain specific data archives. OCLC Systems & Services 23, (1): 35-53.

Karasti, Helena, Karen S. Baker, and Eija Halkola. 2006. Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. Computer Supported Cooperative Work: CSCW: An International Journal 15, (4): 321-58.

Lord, P., and A. Macdonald. 2003. e-science curation report data curation for e-science in the UK: An audit to establish requirements for future curation and provision. JISC Committee for the Support of Research. Online: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf Retrieved 9/30/2008.

Martinez-Uribe, Luiz. 2008. Findings of the scoping study interviews and the research data management workshop. Oxford e-Research Centre.

Treloar, Andrew, David Groenewegen, and Cathrine Harboe-Ree. 2007. The data curation continuum: Managing data objects in institutional repositories. D-Lib Magazine 13(9). Online: http://www.dlib.org/dlib/september07/treloar/09treloar.html. Retrieved 9/30/2008.

Woodbury, P.B., R.W. Howarth, and G. Steinhart. 2008. Understanding Nutrient Cycling and Sediment Sources in the Upper Susquehanna River Basin. Journal of Contemporary Water Research and Education (139): 7-14.

Footnotes

[1] http://cugir.mannlib.cornell.edu/

http://usda.mannlib.cornell.edu/

http://www.w3.org/TR/owl-features/

http://www.fgdc.gov/metadata/csdgm/

http://knb.ecoinformatics.org/software/eml/

http://knb.ecoinformatics.org/

http://ecommons.library.cornell.edu/

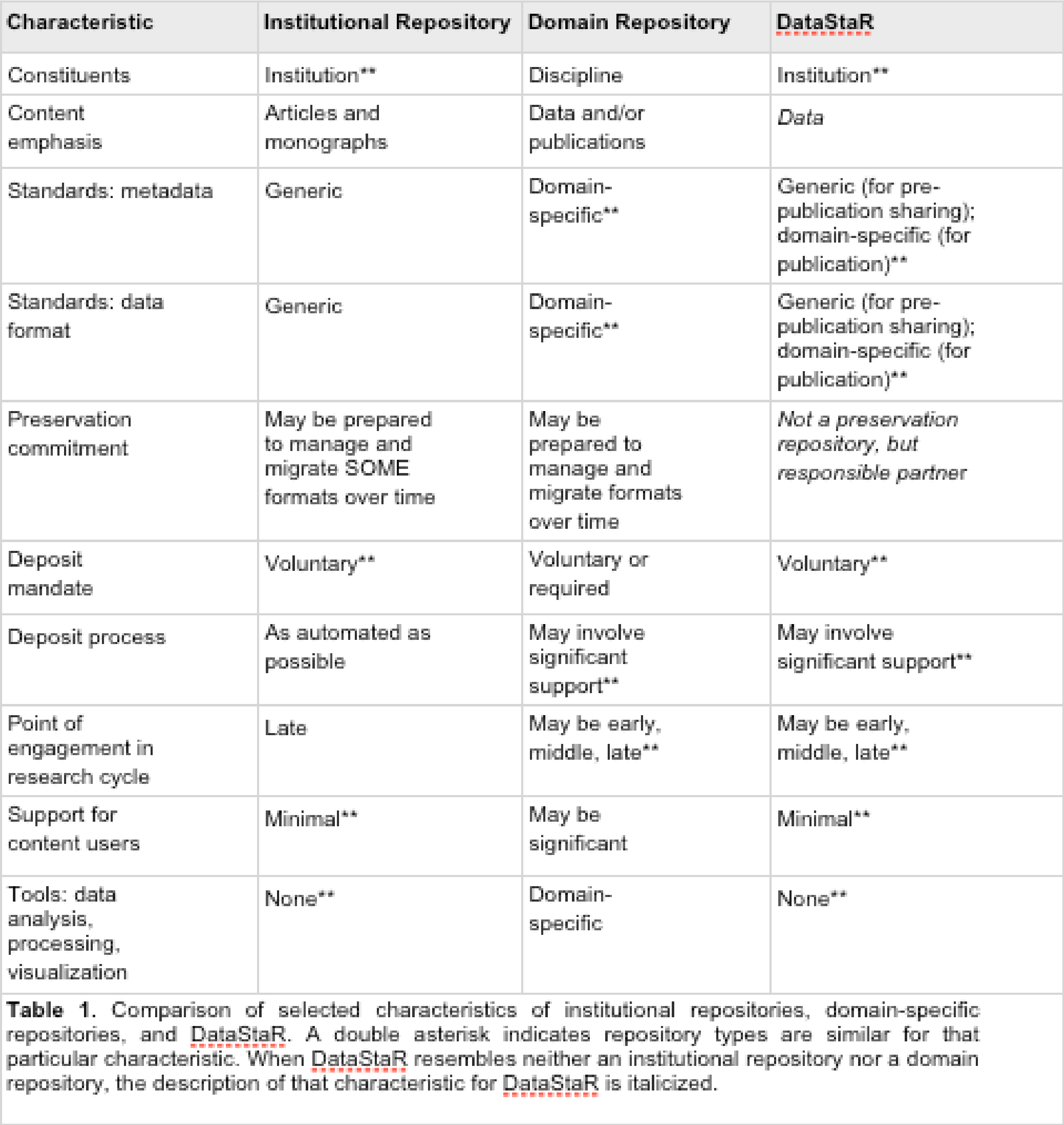
http://vitro.mannlib.cornell.edu/

http://vivo.cornell.edu

http://www.crl.edu/PDF/trac.pdf

http://www.clal.cornell.edu/

http://www.usaep.mannlib.cornell.edu/



Characteristic	Institutional Repository	Domain Repository	DataStaR
Constituents	Institution**	Discipline	Institution**
Content emphasis	Articles and monographs	Data and/or publications	*Data*
Standards: metadata	Generic	Domain-specific**	Generic (for pre-publication sharing); domain-specific (for publication)**
Standards: data format	Generic	Domain-specific**	Generic (for pre-publication sharing); domain-specific (for publication)**
Preservation commitment	May be prepared to manage and migrate SOME formats over time	May be prepared to manage and migrate formats over time	*Not a preservation repository, but responsible partner*
Deposit mandate	Voluntary**	Voluntary or required	Voluntary**
Deposit process	As automated as possible	May involve significant support**	May involve significant support**
Point of engagement in research cycle	Late	May be early, middle, late**	May be early, middle, late**
Support for content users	Minimal**	May be significant	Minimal**
Tools: data analysis, processing, visualization	None**	Domain-specific	None**

Table 1. Comparison of selected characteristics of institutional repositories, domain-specific repositories, and DataStaR. A double asterisk indicates repository types are similar for that particular characteristic. When DataStaR resembles neither an institutional repository nor a domain repository, the description of that characteristic for DataStaR is italicized.