6th International Digital Curation Conference
December 2010

# DataStaR: Using the Semantic Web approach for Data Curation

Huda Khan, Brian Caruso, Jon Corson-Rikert, Dianne Dietrich, Brian Lowe, Gail Steinhart

Albert R. Mann Library,

Cornell University

December 2010

**Abstract**

In disciplines as varied as medicine, social sciences, and economics, data and its analysis are an essential part of researchers' contributions to their respective fields. While sharing research data for review and analysis presents new opportunities for furthering research, capturing this data in digital form and providing the digital infrastructure for sharing data and metadata pose several challenges. This paper reviews the motivations behind and design of the Data Staging Repository (DataStaR) platform that targets specific portions of the research data curation lifecycle (Higgins, 2008): data and metadata capture and sharing prior to publication and publication to permanent archival repositories. The goal of DataStaR is to support both the sharing and publishing of data while at the same time enabling metadata creation without imposing additional overhead for researchers and librarians (Steinhart, 2010). Furthermore, DataStaR is intended to provide cross-disciplinary support by being able to integrate different domain-specific metadata schemas according to researchers' needs. DataStaR's strategy of a usable interface coupled with metadata flexibility allows for a more scaleable solution for data sharing, publication and metadata reuse.

## Overview

Researchers rely on data as scientific evidence of their claims and as the basis for the knowledge that they generate (Arms, 2008). Descriptive metadata allow researchers to define the context needed for future data analysis and further review by themselves and other researchers, and thus adequate metadata are needed for effective data discovery, analysis and reuse. At the same time, the process of metadata creation can require researchers to learn a particular metadata schema or to use specialized tools. Researchers may perceive metadata creation to be too time-consuming and tangential to the overall process of their research and may not learn and use a particular metadata schema unless metadata use is critical or necessary for research (Pritchard, Anand & Carver, 2005). Pritchard, Anand, and Carver (2005) suggest the use of metadata-agnostic repositories and interfaces that automate metadata creation as a means to support metadata use.

Librarians with metadata and/or subject area expertise are in a good position to assist researchers with metadata creation, but, as Steinhart and Lowe (2007) found in their efforts to support research data curation at Cornell University's Albert R. Mann Library, tasking librarians with metadata creation without appropriate tools is not a sustainable approach. Prior to developing DataStaR, Mann Library was engaged in several data curation initiatives, working with faculty and research teams to prepare, describe, and archive scientific data sets (Steinhart & Lowe, 2007). One such initiative involved working with a research group that was studying nutrient and sediment cycling in the Upper Susquehanna River basin. The members of this research group were from multiple institutions and expressed an interest in sharing documents and data within the group prior to publication (Steinhart & Lowe, 2007) as well as sharing their results publicly (Steinhart, 2010). In the process of supporting and training the group to document and publish their data sets using domain-specific metadata, Steinhart and Lowe (2007) realized that the strategy of shifting the bulk of metadata creation to librarians does not scale well with an increasing number of researchers and research groups. In order for more researchers to be able to create metadata without placing unsustainable demands on library staff time, researchers need tools that enable them to do most or all of data documentation themselves with occasional assistance from librarians as needed.

Ann Green and Myron Gutmann's (2007) description of the possibilities for partnerships between institutional and domain repositories further helped crystallize the need for a local, institutionally-based staging repository which enables domain-specific metadata definition before and up to publication (Steinhart, 2010). DataStaR seeks to provide such a service, scaffolding the process of eventual data publication to both institutional and domain-specific repositories (Dietrich, in press). DataStaR, as a staging repository, is not intended to serve as a permanent repository and thus does not itself need to conduct preservation planning per Higgins' digital curation lifecycle model (Higgins, 2008), but the system does address curation of data at different stages of the research process and is designed to support best practices for preservation (Steinhart, Dietrich, & Green 2009).

# DataStaR and the Semantic Web

Semantic Web technologies aim to define and interconnect data in a way similar to how traditional web technologies define and interconnect web pages. In the case of

the traditional web, each web page can be considered a unit of information or entity, and pages are explicitly linked using html links. The Semantic Web also allows data to be shared using "linked data"[1] [2] support where entities can be referenced and their information can be accessed on the web as part of a linked network of data. Entities are identified using URIs or Unique Resource Identifier, similar to URLs, and are described using Resource Description Framework (RDF[3]) statements. These statements describe entities using "<subject> <predicate> <object>" triples where "subject" is the entity, "predicate" refers to a property or relationship for the entity, and "object" can be either a literal value such as text or another URI referencing another entity . Semantic web applications can thus retrieve and integrate this web of statements describing a given entity.

DataStaR's use of semantic web technologies attempts to support more efficient creation of metadata by treating the metadata associated with a particular data set as a collection of statements about that data set, rather than a single, static document. This approach enables the reuse of statements for other data sets, potentially decreasing the effort involved in creating metadata, particularly as a researcher's "collection" of metadata statements in DataStaR grows.

This semantic web approach also enables DataStaR to support metadata creation across multiple discipline-specific metadata schemas. Different metadata schemas are integrated into DataStaR as needed by being converted into semantic ontologies using RDF statements and OWL[4] classes. DataStaR can thus be extended to describe data sets from various disciplines. In addition, DataStaR's use of the semantic web approach enables the reuse of metadata across different metadata schemas through the inclusion of mappings between ontology elements. For example, when DataStaR defines the mapping between Ecological Metadata Langauge (EML[5]) and the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata (FGDC[6]) geographic coverage statements, information entered by the user for a geographic coverage element for an EML data set can be reused for FGDC describing the data set. Researchers can thus use DataStaR to create, share, and publish data sets described by different schemas as required. Furthermore, DataStaR can describe a single data set using multiple metadata schemas when needed.

## DataStaR Application: Architecture and Metadata Creation

Figure 1 provides an overview of DataStaR architecture. DataStaR extends the Vitro software developed by Mann Library at Cornell University and that "combines a Web-based ontology and instance editor with a public display interface" (Lowe, 2009). Vitro is best known as the software underlying the VIVO research networking tool, also developed at Cornell and now expanding to a number of other universities under the sponsorship of the National Institutes of Health (http://vivoweb.org).

DataStaR customizes Vitro[7] to define and specify the relationships between data

[1] Linked Data, http://www.linkeddata.org
[2] Linked Data Design Issues by Tim Berners Lee, http://www.w3.org/DesignIssues/LinkedData.html
[3] RDF, http://www.w3.org/RDF/
[4] OWL Web Ontology Language, http://www.w3.org/TR/owl-features/
[5] Ecological Metadata Language, http://knb.ecoinformatics.org/software/eml/
[6] http://www.fgdc.gov
[7] Vitro, http://vitro.mannlib.cornell.edu

sets, individuals, and organizations. OWL ontologies are used to define the types of entities and what properties or predicates can be used to describe these entities. A data set's metadata input forms are generated based on the associated ontologies. Files uploaded to a data set are stored in the Flexible Extensible Digital Object Repository Architecture or FEDORA[8] repository. DataStaR generates RDF statements to define this file as an entity with a uri and to store file-specific information such as size, content type, checksum, and the unique FEDORA identifier or pid for the FEDORA file.

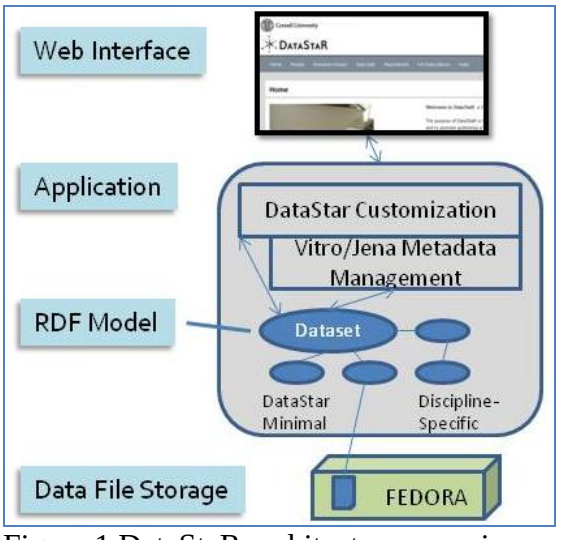


Figure 1 DataStaR architecture overview.

Consider a scenario where a hypothetical environmental scientist named Sara creates a data set using DataStaR. After logging into her account, Sara selects the option to create a new data set and indicates that her intended submission repository is the Knowledge Network for Biocomplexity (KNB[9]) which requires EML. As Figure 2 below shows, the initial data set creation page requires only a few metadata fields, such as title and destination repository, be filled out by the user while the remaining fields, such as data set originator, are automatically generated and assigned to the resulting data set. This core set of DataStaR metadata fields are common to all data sets.

Sara could have selected "to be determined" as the destination repository if she's unsure of her publication plans or if she is only using as DataStaR as a means to share data with authorized colleagues. Intent to publish is not a requirement for researchers to use DataStaR. If no expected publication date is indicated at the time of data set creation, a date one year in the future is included by default. When this date is reached and if the data set has not yet been published, the DataStaR staff may contact the owner or originator of the data set to request an update on the status of the data set. Sara can also define access and modification permissions for different individuals and research groups.

[8] http://http://fedora-commons.org/
[9] Knowledge Network for Biocomplexity, http://knb.ecoinformatics.org/

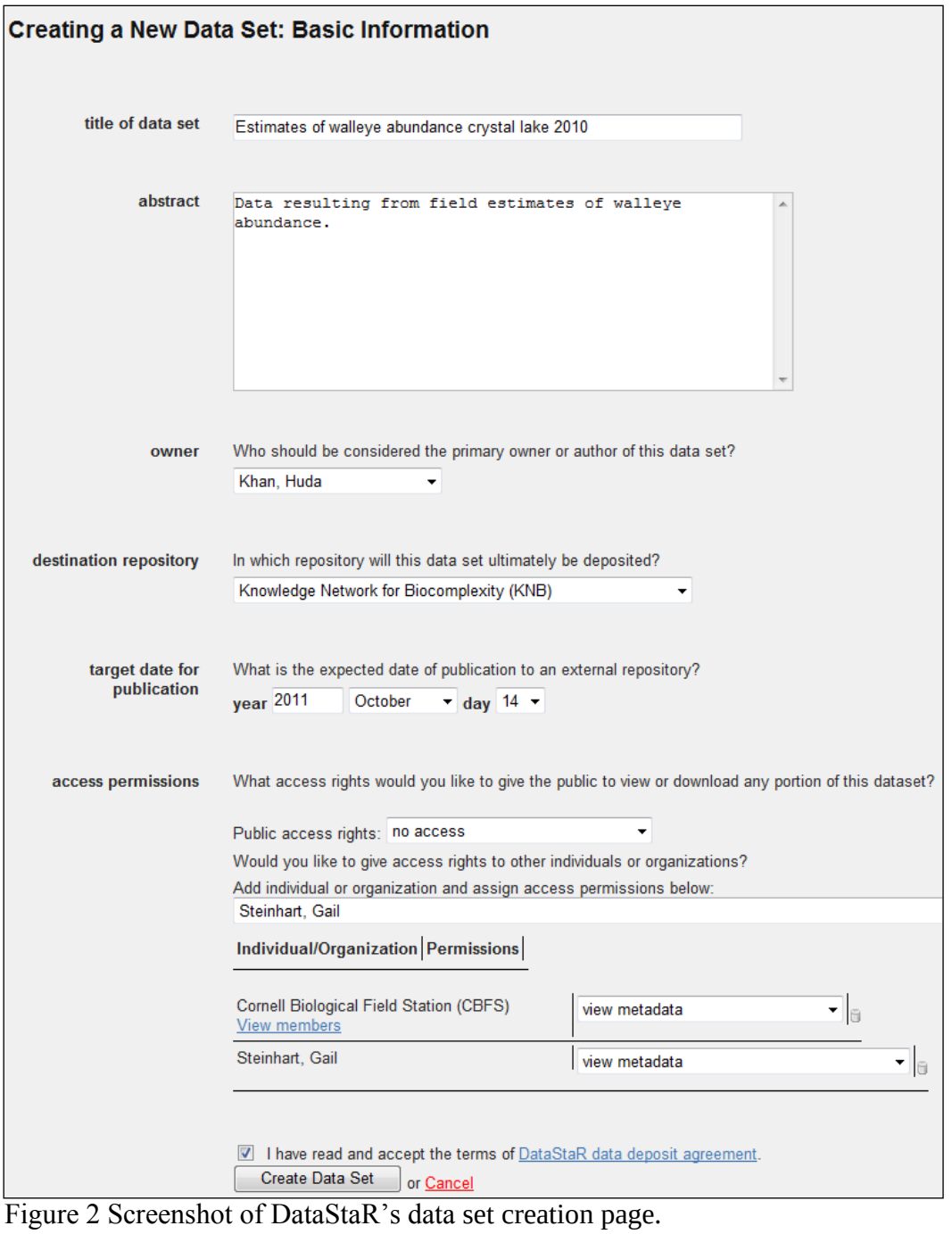


Figure 2 Screenshot of DataStaR's data set creation page.

Within the RDF model, the data set is now defined as an entity with a unique URI with related RDF statements. Figure 3 below provides a simplified sample of these statements as an RDF graph. <dataset> designates the data set uri and <Sara> indicates the owner uri in DataStaR. The "rdf" and "dsr" prefixes designate the RDF and DataStaR specific namespaces respectively.

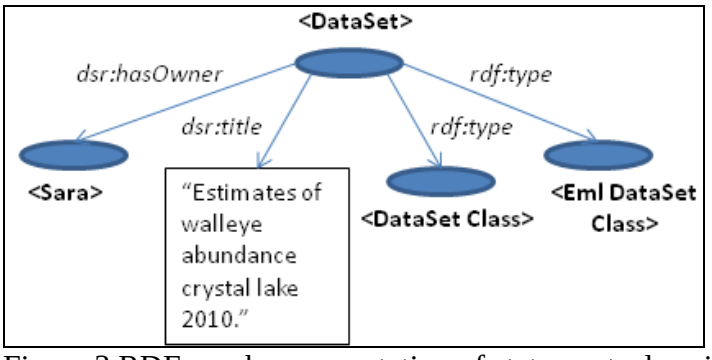


Figure 3 RDF graph representation of statements describing an EML data set.

Once Sara has completed and submitted this form successfully through the interface, she can now view and edit the fields. Because Sara indicated the KNB repository as the destination repository, the system generated a statement defining the data set as having an "EML data set" type in addition to the regular data set type. The EML type triggers the data set view form to include fields and properties that are from the EML ontology. For example, Sara can add geographic coverage information which maps to the EML geographic coverage elements.

Figure 4 shows a high level overview of how these EML statements integrate with the minimal and EML-specific ontologies in DataStaR. The statements shown in the figure are not RDF but simplified versions that show the kinds of information encoded into RDF statements for both statements generated when the data set is created and edited and the ontologies for the core DataStaR and integrated EML schemas.

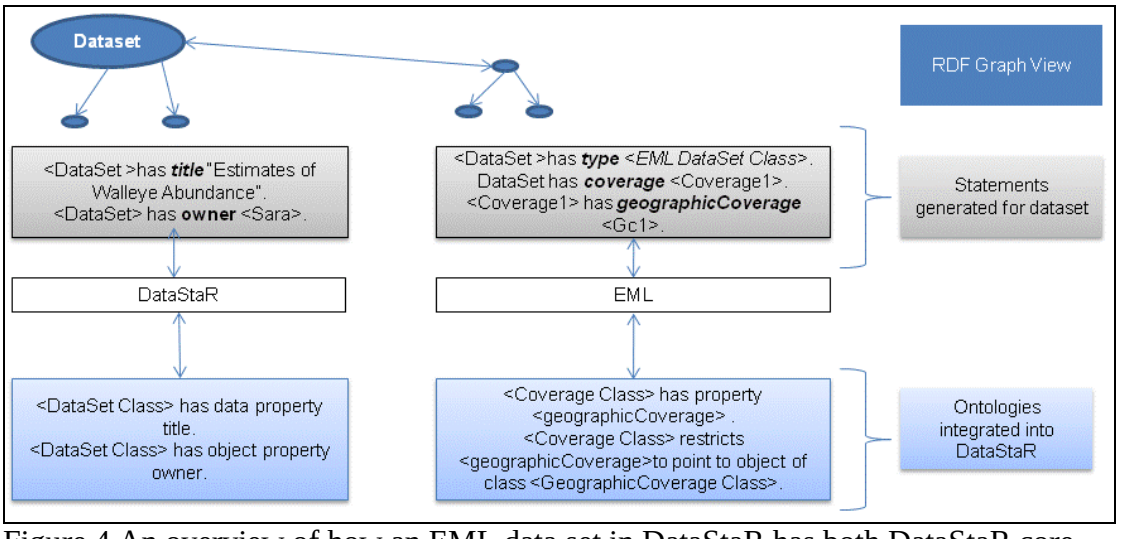


Figure 4 An overview of how an EML data set in DataStaR has both DataStaR core ontology statements as well as EML statements.

Sara continues to edit and share these data sets with colleages or research groups. When her colleagues download the data set, DataStaR returns a zipped file containing the files uploaded as well as separate XML files corresponding to the different schemas with which the data set was associated. In this case, they would receive the uploaded data files, one metadata xml file corresponding to minimal DataStaR metadata and one EML file mapping to the EML statements for the data set. DataStaR creates the EML record using the Gloze application (Battle, 2006)'s transformation of the data set's RDF statements to XML. DataStaR may provide additional changes to

the output from Gloze's transformation for better alignment between the resulting XML and EML specifications. DataStaR uses a similar process to create an EML record when Sara wishes to publish the data set to the KNB repository.

## Challenges and questions

***Named Graphs: Information Integrity and Controlled Access***

One of the appeals of semantic web technologies lies in the potential for linking and integrating data from multiple sources and then being able to query and retrieve information across these different sources. In spite of the desirability of linking data in this manner, a concern that arose during the development of DataStaR was how to maintain information integrity through controlled access while still supporting metadata reuse. If all information in the system is available to all users, it's possible for an individual to edit an entity created and used by someone else (the originator) in such a way that introduces changes or errors into the description of one or more of the originator's data sets. An example we've already encountered has to do with changing roles of research participants. A researcher may be described accurately as the director of a research facility at the time a data set is created, but may later retire, with another individual being promoted to that role. The information in the system is changed to reflect the changes in roles, but it is not necessarily appropriate to change that information for a data set created earlier. We realized it would be necessary to stabilize information about a particular data set to avoid propagating later changes unintentionally.

At the same time, a researcher may wish to give different levels of access to different individuals for the same data set. For example, our example scientist Sara may wish to restrict her data set's public visibility but share her data set with a group of colleagues. She may wish to allow a researcher working on the same project to be able to modify the metadata and she may decide at a certain point in the future that she would like to make the data set visible to the public.

In order to address these scenarios, DataStaR employs private named graphs which are a collection of statements referenceable by a URI. A given data set's information is stored in an associated named private graph. Certain information, such as the title or the graph URI itself, is stored in the public layer of RDF statements while the remaining set of statements for that data set are included in that data set's named graph. Every user can see the publicly accessible RDF statements but access to the named graphs is based on whether or not the user has explicit permissions to view a particular data set.

Consider again the data set created by Sara actually consists of two sets of statements, one set which consists of basic identifying information and another set which is comprised of all other information stored within a named private graph. When Sara created this data set, she specified additional users or groups who could have access to the data or metadata. In accordance with this information, DataStaR created RDF statements defining permissions related to this data set and automatically gave full permissions to the owner Sara. When Sara herself logs in, DataStaR checks for which data sets she has permissions and then adds the corresponding private graphs to the main or "public" graph which is visible to Sara.

If Sara then sees, for example, a set of geographic coordinates (perhaps for a common sampling location) in another data set which she would like to reuse in her data set, she can select it from the list of previously defined coordinates. These coordinates are then copied over into the private graph for the data set she is editing. This copying process also occurs when a data set is first created. The system searches for the object references that are used to describe the dataset and then copies information about these objects into the data set's private graph. For example, our example data set's owner is defined using a statement which declares that the data set has the owner Sara (where Sara is identified by a URI). The system searches for additional statements in the public model describing the URI representing Sara, such as statements describing the label or name associated with the URI, and then copies those statements to the data set's private graph. This copying process allows for the user to see the owner name when they are editing the private graph, whereas without the copying process they would only see the owner uri.

The use of private graphs, though helping to resolve the issue of maintaining information integrity, raises additional questions. When information is copied into the private graph from the public layer or from another data set, that information is a snapshot of the content available at the time of the copy. The question then becomes when information, and what portions of the information, should be synchronized with the public layer, and under what circumstances? For example, data set B may be related to data set A, and data set B's private graph would contain the copy of data set A's title when this relationship was created by the user. If data set A's title changes at some point prior to data set A's publication, data set B would still display the old title by virtue of the information stored in data set B's private graph. This case suggests the need to include a synchronization feature which would allow certain properties to be updated to the information that is present in the public model, if desired, prior to publication or export to another repository.

***Metadata: XML to RDF***

In most data repositories, metadata is stored using XML files based on XML schemas which may allow complex, nested, and ordered elements. In order to be able to integrate different metadata schemas into DataStar, the development team had to consider how to translate the XML Schema Document (XSD) underlying a given XML metadata record into an OWL ontology and how to transform the XML record into a data set described using RDF statements. In addition, the system then should be able to take the resulting data set and transform it back into an XML file consistent with the publication repository's metadata requirements.

Gloze (Battle, 2006) can help to convert a metadata specification's XSD into a set of OWL classes of objects and corresponding predicates. In the case of very complex metadata schemas, we may include selective portions of the schema, for example only those elements available in commonly used metadata creation or editing tools for that schema, or the most commonly used elements of a particular schema. The ontology resulting from Gloze can be refined or extended as needed. The DataStaR team has explored the integration of EML as well as the custom schema employed by the Virtual Center for Language Acquisition (VCLA[10]) to store metadata for a linguistic

[10] Virtual Center for Language Acquisition, http://vcla.clal.cornell.edu/

study. The integration of additional metadata schemas has exposed certain challenges in the conversion of XML to equivalent RDF statements and in the displaying of these statements in a way which makes sense to those editing the statements through the DataStaR interface. These challenges include (a) converting implicitly ordered XML elements in a parent element and (b) generating an interface for XML schema restrictions involving "choice" where only one element out of a set of options should be included in a parent element.

***Nested repeatable XML elements and implicit order***

In some cases, XML files have an implicit ordering that then needs to be correctly captured in the RDF statements. For example, an EML record can contain multiple method steps nested in the methods element. Although there is no explicit order number given to these elements, the elements are listed in a specific order. When configured to order these nested elements, Gloze generates an RDF sequence element which describes the order of nested elements using predicates such as "rdf:_1" and "rdf:_2". In order to be able to use these predicates, DataStaR's ontology would have to create a separate "rdf:_x" predicate, where "x" corresponds to a number, for an entire range of numbers i.e. a separate property for rdf:_1, rdf:_2, rdf:_3 etc. This solution would either result in a very large number of "rdf:_" properties or the need to add a new "rdf:_" property every time a new order number was needed. DataStaR adopted a different solution, indicating order by specifying a set of intermediate entities that link the parent object to the child object while providing ordering information. As part of integrating EML into DataStar, special "ordering" objects and properties were defined in the DataStaR ontology. These properties can be extended based on the type of objects being ordered. Figure 5 shows the mapping from the XML method steps to the generic RDF ordering relationship as well as the extended relationship "orderedMethodStep".

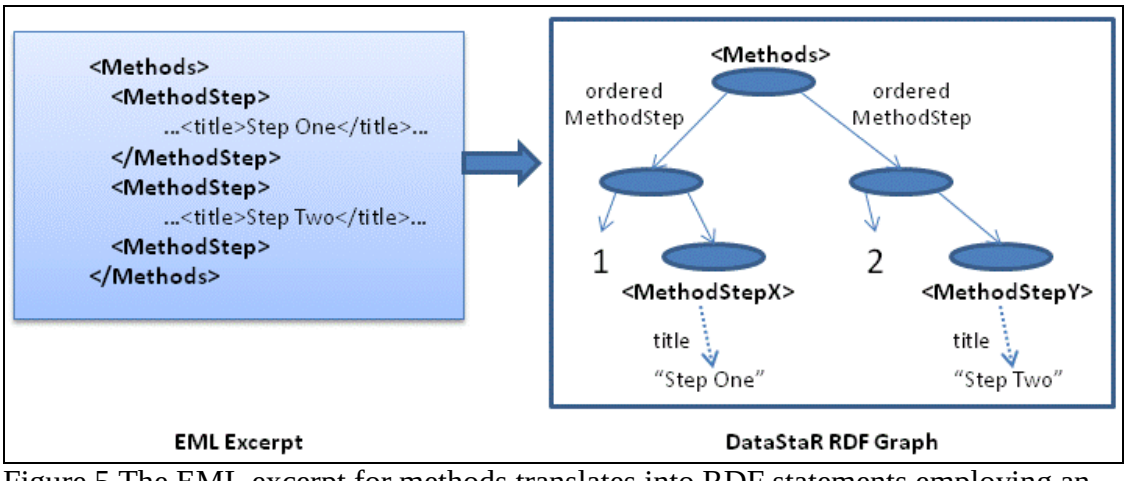


Figure 5 The EML excerpt for methods translates into RDF statements employing an intermediate ordering context. Ellipses in the EML and dotted arrows in the RDF graph representation indicate additional child elements or hierarchy of RDF statements respectively.

The DataStar interface has to then recognize these ordering constructs in addition to the base ontology for the schema. We modified the data set view page to order content with respect to the order values for these intermediate objects, and updated the metadata field editing page to allow for the addition of new elements while, in the back

end, creating new intermediate objects to define their order as the last in the sequence. For example, when Sara edits the "methods" field for an EML data set and adds a new method step when two method steps already exist, the new method step will be interpreted and displayed as third in the sequence of method steps. We expect to keep updating the interface to allow for a more seamless way of ordering these elements on the same page without having to submit or refresh the page itself.

***XML Choice: Which options to display?***

XML schemas use the "choice" element to specify that an element can contain one and only one of multiple kinds of nested elements. As an example, consider that in an EML record, a "TextType" element, which is used to contain text, may consist of *either* a "section" element *or* a "para" element (short for paragraph). OWL, while capable of expressing the minimum or maximum number of section and para elements allowed, does not have a direct equivalent to XML's choice element. If Sara, our example scientist, were to use DataStaR using this ontology, she would see that, where a TextType entity is included such as in the case of a MethodStep, she can edit two text areas, one entitled "section" and one entitled "para". The interface would not indicate that she only needs to fill out one input. Currently, DataStaR reviews these situations on a case by case basis, updating the integrated ontology to include choices that are consistent with EML but that don't include more options on the interface than necessary. For example, in the case of method steps for an EML method field, we restricted the ontology to include only a section as being part of a method step, allowing the interface to display just the inputs for section. Future work will explore an ontology-based approach to resolve this issue, such as the use of additional annotation properties to describe which field out of different choices should be selected given which element is being edited.

## Current Status and Ongoing Work

The first production version of DataStaR is intended to be ready for use in early 2011. We have developed several partnerships with research teams that intend to use DataStaR to store and publish data sets and that include: Agriculture, Energy and the Environment Program (AEEP), Cayuga Lake Watershed Network, Cornell Biological Field Station, Cornell Plantations Natural Areas Program, the Loon Project, the Virtual Center for Language Acquisition, and the Data Conservancy project (Steinhart, 2010). Table 1 below shows the publication repositories and their metadata specifications that DataStaR will support.

| Repository | Metadata Specification |
|---|---|
| Cornell University Geospatial Information Repository (CUGIR[11]) | Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC-CSDGM) |
| eCommons@cornell[12] | Modified Dublin Core |
| Knowledge Network for BioComplexity (KNB) | Ecological Metadata Language (EML) |
| Data Conservancy | TBD |

Table 1. Publication repositories for data sets in DataStaR and corresponding metadata specifications.

[11] CUGIR, http://cugir.mannlib.cornell.edu
[12] eCommons@Cornell, http://eCommons.cornell.edu

The DataStaR development team has identified several important interface usability issues such as the ability to add multiple elements on a page without having to refresh or reload the page. For example, the user should be able to add multiple method steps and order them for an EML methods section and they should be able to add multiple keywords while editing the keyword set for a data set. The development team is exploring future opportunities for conducting usability testing on the interface employing faculty or graduate students that are representative of the researchers who would use DataStaR. Furthermore, we will also need to explore how researchers from different domains may have different requirements or workflows and what subset of these requirements we will be able to support using a single system.

Another area for ongoing development is supporting the different workflows for different repositories. In addition to the specific metadata standards mentioned in Table 1, DataStaR will integrate support for data set publication to repositories, such as the Data Conservancy, which support or plan to support the Simple Web-service Offering Repository Deposit (SWORD[13]) protocol. For some repositories such as KNB, the Data Conservancy, or eCommons, a direct submission from DataStaR on publication may be possible. For other repositories with unique architecture or submission procedures, DataStaR may have to create a submission package that would then need to be submitted manually. In addition, the current interface only allows end-users to select a single publication repository and, in the case of KnB, generate EML fields in the resulting data set. DataStaR will also need to support cases allowing for more flexibility, for example if a person wishes to submit a data set to eCommons, a repository that requires modified Dublin Core metadata, along with a discipline-specific metadata record (as a supplementary document).

As work proceeds with DataStaR, we have seen an increased interest in the application. Some researchers with whom we have worked previously intend to use DataStaR as part of their data dissemination plans. Other institutions and projects are exploring the use and adaptation of Vitro, a core component of DataStaR, alone or in combination with the VIVO[14] research networking ontology. One such project is the Australian National Data Service (ANDS[15]) which is developing a national data registry and has funded enhancements to Vitro as a metadata acquisition and submission tool at several participating Australian universities including Queensland University of Technology, Griffith, and the University of Melbourne. We continue to explore the integration of additional metadata standards and the improvements to the design of the interface to support researchers in their metadata creation for research data.

## Acknowledgments

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 and do not necessarily reflect the views of the National Science Foundation.

[13] SWORD, http://swordapp.org
[14] VIVO, http://vivoweb.org
[15] ANDS, http://ands.org.edu/funded/eif-fast-start.html

# References

[Internet journal]Arms, W.Y. (2008). Cyberscholarship: High performance computing meets digital libraries. *The Journal of Electronic Publishing 11,(1).* Retrieved August 6, 2010, from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0011.103

[proceedings] Battle, S. (2006). Gloze: XML to RDF and back again. *Proceedings of First Jena User Conference, Bristol, UK.*

[journal article] Dietrich, D. (In press). Metadata management in a data staging repository. *Journal of Library Metadata*.

[journal article] Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems & Services, 23*(1), 35-53.

[journal article] Higgins, S. (2008). The DCC Curation Lifecycle Model. *The International Journal of Digital Curation 3,(1).* UK: UKOLN, University of Bath.Retrieved August 6, 2010 from http://www.ijdc.net/index.php/ijdc/article/download/69/48.

[proceedings] Lowe, B. (2009). DataStaR: Bridging XML and OWL in Science Metadata Management. *Metadata and Semantic Research Third International Conference, MTSR 2009. Milan, Italy*: Berlin: Springer.

[report] Pritchard, SM., Anand, S. & Carver, L. (2005). *Informatics and knowledge management for faculty research data.* Retrieved August 6,2010 from http://net.educause.edu/ir/library/pdf/ERB0502.pdf.

[proceedings] Steinhart, G. & Lowe, B. (2007). Data curation and distribution in support of cornell university's upper susquehanna agricultural ecology program. *DigCCurr2007, an International Symposium on Digital Curation. Chapel Hill, NC.*

[journal article] Steinhart, G., Dietrich, D., & Green, A. (2009). Establishing trust in a chain of preservation: The TRAC checklist applied to a data staging repository (DataStaR). *D-Lib Magazine Magazine, 15*(9/10).

[proceedings] Steinhart, G. (2010). DataStaR: a data staging repository to support the sharing and publication of research data. *International Association of Scientific and Technological University Libraries, 31st Annual Conference. West Lafayette, IN.*