Discovery Informatics: AI Takes a Science-Centered View on Big Data
AAAI Technical Report FS-13-01

MetaShare: From Data Management Plans to Knowledge-Based Systems

Leonardo Salayandia Deana Pennington Ann Q. Gates Francisco Osuna
Cyber-ShARE Center of Excellence, University of Texas at El Paso
500 W. University Drive
El Paso, Texas 79968, USA

Abstract

MetaShare is a knowledge-based system that supports the creation of data management plans and provides the functionality to support researchers as they implement those plans. MetaShare is a community-based, user-driven system that is being designed around the parallels of the scientific data life cycle and the development cycle of knowledge-based systems. MetaShare will provide recommendations and guidance to researchers based on the practices and decisions of similar projects. Using formal knowledge representation in the form of ontologies and rules, the system will be able to generate data collection, dissemination, and management tools to facilitate tasks with respect to using and sharing scientific data. MetaShare, which is initially targeting the research community at the University of Texas at El Paso, is being developed on a Web platform, using Semantic Web technologies. This paper presents a roadmap for the development of MetaShare, justifying the functionality and implementation decisions. In addition, the paper presents an argument concerning the return on investment for researchers and the planned evaluation for the system.

Introduction

The MetaShare knowledge-based system is being developed to aid researchers in defining data management plans that include formal semantic metadata to support the data management life cycle. MetaShare leverages the development cycle of knowledge-based systems and the research data life cycle to promote research data discovery. Data management plans produced with MetaShare are intended to be actionable (Salayandia, Gates, and Pennington 2013), while other such systems focus on guiding researchers to produce plans that present data and management procedures for project evaluators. By sharing data management knowledge among researchers of a community, MetaShare can assist researchers in making decisions on their plans based on established practices and decisions that others have done in the past. A feature of MetaShare is to provide a recommender system that can be used by researchers to plan and implement data stewardship activities, while simultaneously building a knowledge base about the data. By leveraging researcher activities to collect and steward data, MetaShare represents this knowledge in a machine-parsable form and connects associated data accordingly. This results in a tight coupling between knowledge and data, a key challenge of Discovery Informatics (Gil and Hirsh 2012).

The benefits of MetaShare are as follows:

- **Scientific data management activities are automated.** Data stewardship requirements, such as dissemination, archiving, and definition of security policies are automated through rule-based systems. Such rules are generated from decisions made by researchers during data management planning and implementation of data stewardship requirements in designated storage facilities.
- **Scientific data are discoverable and sharable.** Based on decisions made by researchers during data management planning, tools are generated to facilitate data collection and ingestion into knowledge-based systems to support querying and inferencing services.
- **Data management recommendations are provided.** By continuously collecting decision information about data management planning from the scientific community, the dynamic knowledge-based system can provide recommendations based on the decisions that colleagues with similar needs have made on past projects.

The paper is organized into six sections that present the following: a discussion of the links between the scientific data life cycle and the development cycle of knowledge-based systems and how MetaShare is being designed to exploit the linkages to provide a user-driven system; a description of the type of knowledge that is being captured by MetaShare and the functionality that it supports; a description of the mechanisms for converting the knowledge captured in MetaShare into useful tools for researchers to collect and process their data; a discussion of the usage and management of data supported by MetaShare for accessing, using, and managing data; a plan for evaluation, and concluding remarks.

Connection of the Data Lifecycle and Knowledge-Based Systems

Research funding agencies, such as the National Science Foundation, recently began enforcing policies that require

researchers to create data management plans. A data management plan describes the types of data to be collected and management activities required for adequate data stewardship. Such policies have had the initial effect of bootstrapping the process of researchers documenting proper data stewardship practices; however, many research projects lack the support to implement fully proposed data management plans because of the effort typically required. A key barrier in achieving data stewardship is the challenge of defining adequate metadata due to a well recognized issue– disparity in work and benefit (Grudin 1994). Researchers possessing the knowledge to document the data, e.g., to describe the data with metadata, may not recognize the return on investment (ROI) for their efforts. They are understandably unwilling to engage in cumbersome activities that may not necessarily benefit their own research. A key MetaShare strategy is to leverage the knowledge that researchers invest in required data management planning to construct a knowledge base that simplifies metadata creation when data are collected. As depicted in Figure 1, MetaShare is being designed around researcher activities that relate to both the data life cycle and the development cycle of knowledge-based systems.

During the first phase in the scientific data life cycle – *planning* – researchers articulate what data they will be collecting, what metadata standards they will use, how they intend to manage it, and when it will be shared with others. Similarly, the initial phases of knowledge-based system development are *identification*, *conceptualization*, and *formalization*, where knowledge engineers and domain experts scope the domain of discourse, create formal knowledge representations, and identify adequate problem solving methods. By scoping knowledge-based systems to a particular type of system that supports researchers in managing their data, researcher activities from the planning phase can be leveraged to drive the initial development phases of a knowledge-based system. For example, identifying the type of data to be collected serves as a basis for constructing an ontology that formally describes the data. Procedural knowledge in the form of rules can be defined in advanced for common data policies, allowing researchers to be guided through a decision tree to choose data policy rules that are adequate for their project and that are customized to execute in their particular data storage environment.

The *collection* phase of the data life cycle is where data generated from scientific activities are recorded. The researcher uses tools to support the collection of data, which may range from lab notebooks and software tools, to sophisticated instrumentation that combines the activities related to the scientific experiment and the activities of collecting data. From the point of view of knowledge-based system development, the *implementation* phase is where the knowledge base is constructed, which involves the acquisition of knowledge according to the structure provided by the knowledge representation formulated in the previous phase. Following an approach similar to that of the Protégé knowledge-based system development environment (Gennari et al. 2002), MetaShare supports the generation of knowledge acquisition tools based on ontologies created by users. This approach is intended to alleviate the issue of disparity of work and benefit mentioned earlier, where the researcher generates data documentation for somebody else to use. An ROI for the researcher in using MetaShare to create a data management plan is that data and metadata collection tools can be automatically generated for researchers to use in their work. This functionality of MetaShare is described in more detail below.

Once data has been collected, the next phases in the data life cycle involve *analysis and use* of data, *disseminating and sharing* data, and eventually, *archiving* data. Researchers use many tools to analyze and use data. Often, researchers are required to manually reformat the data to be able to use it across multiple tools. Similarly, researchers have to prepare data for dissemination and sharing, i.e., data is reformatted and documented by attaching metadata that provides a context for a particular audience. Finally, researchers have to determine what subsets of the collected data are worthwhile archiving for the longer term, taking into account storage and preservation costs. From the point of view of knowledge-based systems, researcher activities related to using and managing data can be facilitated or automated, providing an additional ROI on the researcher's initial efforts to create a management plan in MetaShare. For example, querying mechanisms can be used to extract data from the knowledge-based system, and converters can be used to transform the data to appropriate formats for use. The knowledge-based system can be used to infer related metadata that can be used to reformat data for dissemination. Lastly, procedural knowledge in the form of rules can be used to automate management-related activities.

The MetaShare vision is to formalize knowledge used by researchers during data management planning in order to generate a scoped, data-centric knowledge base, to use that knowledge to generate data collection instruments that ingest data into the knowledge base, to utilize the inference mechanisms and generic problem-solving methods of the knowledge-based system to prepare data for use and sharing, and to automate data management activities. The functionality of MetaShare's knowledge-based system is being made available through the Web, i.e., utilizing Semantic Web technologies.

Knowledge Scoping Process

As described earlier, the type of knowledge captured in the MetaShare knowledge-based system is intended to support two main functions: 1) provide recommendations to researchers creating data management plans based on practices and decisions from a community that has used the system for similar purposes, and 2) generate tools to collect and manage data and metadata based on researcher's descriptions of data and decisions made in the planning phase.

The MetaShare upper-level ontology, illustrated in Figure 2, provides structure to organize information related to data management plans. Information includes research project, type of data involved, people involved and their disciplinary background, related organizations, and tools.

The generic *data* concept in the MetaShare upper-level ontology is intended to be subclassed by researchers; they either create their own classes or import classes from existing

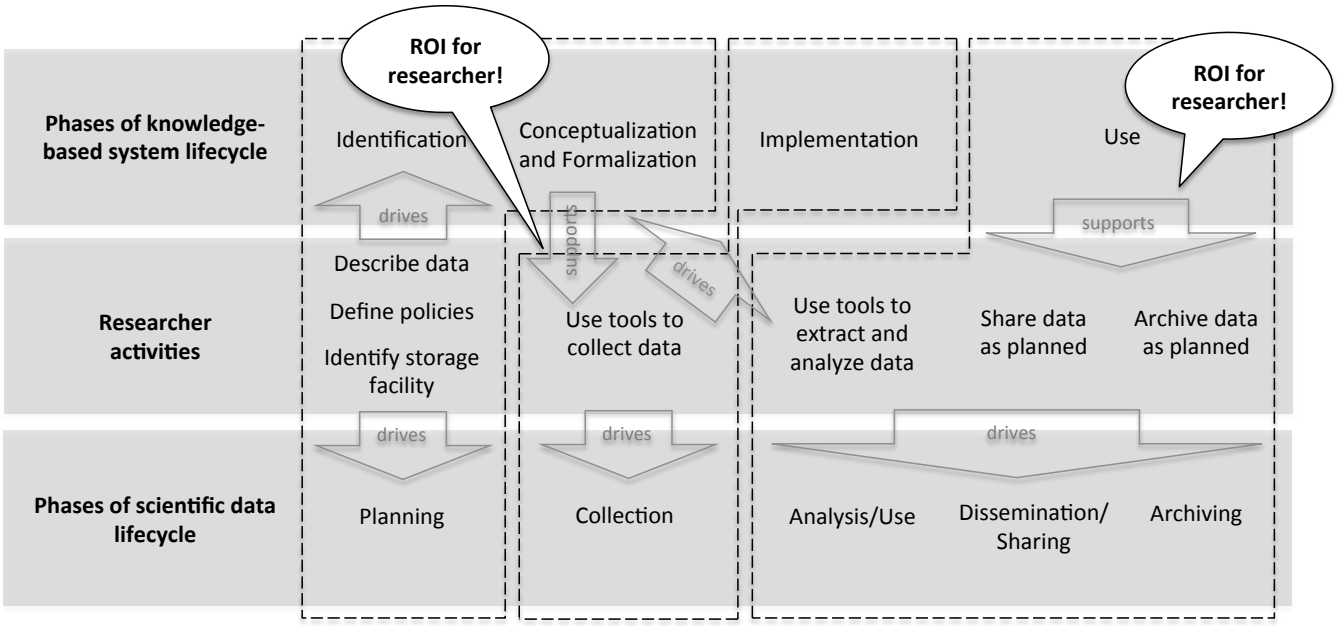

Figure 1: A relation of researcher activities to the cycles of a knowledge-based system and scientific data.

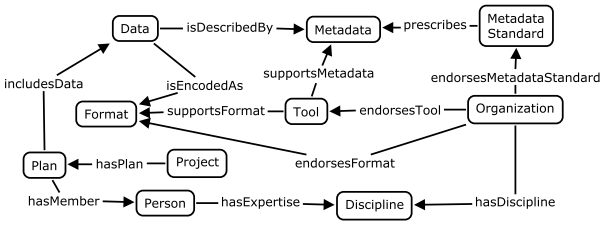

Figure 2: Concept map of MetaShare upper-level ontology.

ontologies as a result of their activities to describe their data in the data management planning phase. Data descriptions can initially be high-level, only identifying keywords. The ontology created at this point would be named subclasses of *data*. As the project progresses from initial planning phases to data collection phases, data descriptions can be refined incrementally, adding properties to the initial data subclasses. Focusing initially on structured data, additional data descriptions that can be documented through MetaShare include the description of dataset variables, which would also include units of measurement, format on which the dataset will be encoded, and metadata standards that are applicable to the dataset.

Based on the ontology, data descriptions that researchers provide for their data and projects can be leveraged to provide recommendations for others. If a researcher is a new user of MetaShare, then he or she enters profile information, including disciplinary background. In the future, MetaShare will use information from external frameworks that maintain expertise information about researchers, e.g., VIVO (Börner et al. 2012), allowing the system to relate *disciplines* to *organizations* that endorse *tools* and *metadata standards*. For example, the Incorporated Research Institutions for Seismology (IRIS, http://www.iris.edu) is an organization that provides various software tools and metadata standards for seismic data. Determining that a researcher has expertise in seismology would lead to recommendations about the types of data that may be applicable based on the tools and metadata standards supported by IRIS. Based on user profile information, MetaShare can provide recommendations by exploiting the content of individual data management plans in a non-consumptive research fashion, i.e., where computational analysis is used on aggregate content and that a researcher does not see directly (Borghi and Karapapa 2011). For example, Figure 3 shows a text field for researchers to enter a *keyword* that describes their data. As the researcher starts to type a keyword, several keywords are recommended based on syntax match, as well as based on keywords that other researchers with similar backgrounds have used. Similarity of researcher backgrounds is determined by their disciplinary backgrounds or the types of tools they use. The researcher entering the keyword has the option to choose one of the recommended keywords, in which case, he or she would be effectively reusing an ontological concept previously defined by another researcher. Alternatively, the researcher may choose to keep typing a new keyword, which would result in a new ontological concept being added to the MetaShare ontology. Hence, the MetaShare interface provides a way for researchers to contribute to the community-built ontology of MetaShare or to reuse concepts from it without coding or training with respect to knowledge representation technologies.

Typically, data management plans are limited in scope to the end products of a research project. However, processes by which scientific data are collected and progres-

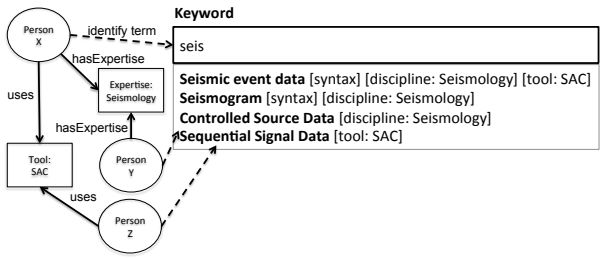

Figure 3: User interface where keywords are recommended based on syntax matches and researcher similarity.

sively transformed are critical to document and seldom included in data management plans. The capture and use of such documentation has been addressed by efforts related to data provenance (Belhajjame et al. 2012; Simmhan, Plale, and Gannon 2005). By providing the option to import data descriptions from external ontologies, the process by which data is transformed can also be included. For example, Workflow-Driven Ontologies can be used to describe the process behind the creation of a data product (Salayandia 2012; Salayandia, Gates, and Pinheiro 2012).

In addition to ontologies, the MetaShare knowledge-based system includes procedural knowledge in the form of rules. Based on decisions that researchers make with respect to dissemination and archive policies, the system selects rules to automate data management tasks. Rules can be executed by frameworks, such as iRODS (Moore 2008). Figure 4 shows a decision diagram that captures the process supported by MetaShare to identify data management rules about data confidentiality. Assuming initial use cases of managing structured data, the current decision branches include the application of a rule to make a dataset confidential by identifying the columns of the dataset that have confidential information, and either altering the values to make the dataset records effectively anonymous, or to filter out the columns altogether. The alternative of describing the confidentiality procedure in text form is also available, which can be used to document more complex procedures; such procedures would not be implemented as rules. However, the expectation is for MetaShare to be an extensible system, where data management rules that are implemented for a given project can be made available for others, discoverable and accessible through the decision process interface.

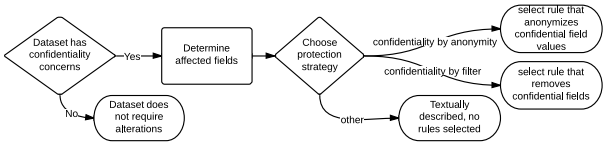

Figure 4: Decision diagram to determine confidentiality rule to apply to dataset.

In iRODS, for example, a rule that implements the confidentiality policy to filter out the third column of a dataset of type *X* could be as follows:

```
confidentiality_policy { on ($dataset_type == "X") {
msi_filter_column ($dataset, 3); } }
```

The iRODS framework includes a rule engine that determines whether to execute a rule based on the condition portion of the rule, i.e., the *on (...)* portion of the rule. The main body of the rule determines the work to be done once the rule is executed, and it typically consists of a series of function calls, or micro service interface (msi) calls. Msi calls are C-functions that a data manager can code to implement data management routines. In the case of MetaShare, a set of msi functions would be pre-defined, which could be used to automatically construct rules that match the decisions made by researchers with respect to data policies and the types of datasets of their projects.

Similar to choosing rules that implement data confidentiality policies, researchers choose rules to implement data policies for dissemination and archiving. Researchers are guided through decision processes to choose when data will be made available to the public, where and how it will be made available, and which parts of the data will be archived for the long term.

Binding Data to Knowledge through Automatically Generated Tools

As previously mentioned, MetaShare leverages researcher activities in the data management planning phases to create ontologies that describe the data of a given project. The ROI for researchers describing their data is two-fold: 1) the resulting ontologies are applied to generate tools that support their data collection efforts, and 2) knowledge bases are implemented around the data to support their activities with respect to use and dissemination of data.

For example, an ongoing biology research project at the University of Texas at El Paso requires collection of gene sequence data that are geographically dispersed. The collected data will be ultimately submitted to the GenBank DNA sequence repository, which requires specific data format and metadata. However, use and analysis of the data by the researchers involved in the project require data to be converted into various formats in order to accommodate their existing toolset. The present practice of data collection involves the use of various spreadsheet and text files. Data collection activities through these files are cumbersome and error-prone, since consistency among the files is manually maintained. While the structure of the current data files was designed by the researchers to support their own processes, it will require considerable effort to disseminate the data beyond the project.

In an initial meeting with researchers of this project, the MetaShare team identified Web forms, which are generated from an ontology, as a viable alternative to the researchers' current data collection method. In accordance to the researchers' description of their data, Figure 5 shows a generated Web form to collect *DNA sequence* data. This type of data has the variables of *sequence string*, *metadata*, *identifier*, *organism*, and *length*. By using the Web form, researchers not only facilitate their requirements of collecting data as a geographically-dispersed team, but submitting data

through the form results in data that is semantically annotated with respect to the ontology. The outcome is a populated knowledge base that links the data to its semantic description.

Researchers may also reuse existing ontologies and keyword suggestions inferred by MetaShare within a specific domain type as the knowledge base is primed. The knowledge base yields an environment in which transformers can be applied in order to generate various formats of the data collected to aid research activities.

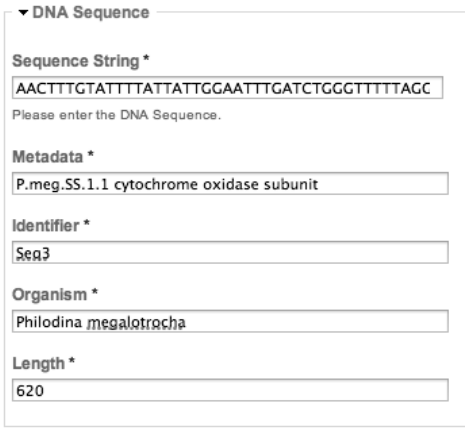

Figure 5: Web form to collect DNA sequencing data, generated from a MetaShare ontology.

As projects and collaborators are identified, generators can be developed from the ontologies. Given a finite set S = $\{x_1,x_2,...,x_n\}$ where $[x_1,x_2,...,x_n]$ are data types provided by researchers, there is a one-to-one correspondence with set F = $\{y_1,y_2,...,y_n\}$ where $[y_1,y_2,...,y_n]$ are the exact pairings or bindings to set S of data collection instruments that ingest data into the knowledge base. These bindings can then be employed to generate Web forms that capture the knowledge to be consumed as represented in Figure 5.

Using and Managing Data through Knowledge-Based Systems

Generated collection instruments are associated with metadata derived from the previously-created data management plan. Such metadata is used to automatically annotate the data in the knowledge acquisition process. In MetaShare, this includes project and investigator information, and the formats and semantics of the data to be collected. Hence, the semantics, which have been developed by a community through use of MetaShare during data management planning, become embedded into automatically generated metadata when data collection begins. Additionally, because the collection instruments are automatically generated, metadata about the logical organization of the data can be automatically generated. These metadata, while not totally comprehensive in all cases, are a significant improvement over existing manual methods and are expected to be adequate for many uses, including:

- **Extracting data with appropriate metadata standards and formats.** Data can be used in multiple ways, both by the original investigator and by others. Depending on the intended use by analysis and other tools, data may need to be extracted to comply with different formats or with different metadata. MetaShare will enable export of data into commonly used formats (e.g., CSV and RDF) and metadata standards (e.g., ISO and FGDC). Additionally, data and metadata can be contributed more easily to emerging registries, such as those being developed by the National Science Foundation DataNet initiative. These registries provide single-point queries for distributed data by either humans or machines.
- **Discovering related data.** The MetaShare upper-level ontology provides a structure that can be used to discover related data. For example, datasets can be found based on common data descriptions, common tools used, or common disciplinary backgrounds of researchers. The MetaShare upper-level ontology can be extended or aligned to other ontologies in order to provide discovery functionality for other scenarios. Standardized Semantic Web technologies were employed in order to maximize compatibility with other ontologies.
- **Supporting data integration efforts.** There are numerous approaches to automatic data integration, all of which depend on adequate, machine-readable metadata. In particular, emerging semantic approaches for data or data/model integration will be enabled through MetaShare. One such approach, Semantic Service Orchestration (Wilkinson et al. 2010), is being developed in conjunction with MetaShare and will provide integration functionality (Del Rio et al. 2013). Semantic Service Orchestration uses the semantic descriptions of service input and output data types to determine intermediate service steps that can transform a given source dataset into the format and structure required by a target service, removing the need for researchers to manually transform data.
- **Automating data management policies.** Although data management plans are required, there is little support for scientists to implement the policies and procedures specified in the plan. MetaShare is being developed to interoperate with iRODS (Moore 2008). Plans developed in MetaShare are coupled with iRODS rules, which are developed and reused among the research community to automatically implement data policies and stewardship activities. For example, a data management plan may specify that all data will be shared 24 months after it is collected. iRODS can implement the rule that automatically triggers an operation to change permissions on data at 24 months after the original creation timestamp.

Plan for evaluation

MetaShare will have to go through several iterations of development, testing, and user feedback before it can be adopted at a wider scale. The initial version is targeting researcher projects at the University of Texas at El Paso, for which MetaShare's knowledge base is being primed with

terms from established hydrology and ecology communities (Maidment 2013; Porter 2013). Additionally, Web forms are being targeted as the initial type of knowledge acquisition tool to be generated from MetaShare, as they are specifically required by our current researcher collaborators.

While MetaShare is currently in the initial iteration of development, there is a plan to evaluate its effectiveness to support researchers in their data stewardship activities.

- **Determine relevant recommendations.** As mentioned above, recommendations to researchers are based on decisions that others with similar backgrounds have done in the past. However, determining the right type of similarity is part of the research to be conducted. User interaction information can be used to conduct quantitative analyses to determine the ontological relations that lead to relevant recommendations. Additionally, a ranking of recommendation options could also be determined.
- **Determine a core set of data policy rules.** As previously mentioned, MetaShare will provide predefined data policy rules from which researchers can choose by following decision trees as they construct their data management plans. iRODS provide a highly-customizable environment that can be used to implement such rules. While it is not feasible to provide data policy rules for all projects and types of data, a library of common rules is envisioned that can be reused by researchers to manage their data. Following the Pareto principle, user feedback should determine a core set of data policy rules that can be used to manage a large number of projects and datasets.

Conclusion

MetaShare is being designed to create actionable data management plans to support researchers in their data sterwardship activities. It provides recommendations during planning, automated generation of data collection tools, and automated generation of substantial metadata. These benefits, which are directed towards data-related tasks that researchers need to accomplish, provide motivation for researchers to devote the time needed to describe data and policies. Although there may be an investment in effort, MetaShare maintains and generates knowledge that allows researchers to more effectively share, discover, and reuse research data. It is a practical approach to bind knowledge more closely to data, a key element of discovery informatics (Gil and Hirsh 2012).

MetaShare developers have been using structured dataset use cases and developing knowledge bases with data descriptions from the UTEP research community, who are serving as beta testers. These early adopters are validating the recommendations provided by MetaShare and evaluating its effectiveness. Additional evaluation efforts are in the works to scale MetaShare to a wider audience.

Acknowledgments

This work is supported in part by the National Science Foundation (NSF) under CREST Grants HRD-0734825, HRD-1242122 and CI-Team Grant OCI-1135525. 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 NSF.

References

Belhajjame, K.; Deus, H.; Garijo, D.; Klyne, G.; Missier, P.; Soiland-Reyes, S.; and Zednik, S. 2012. Prov model primer. Technical report, W3C.

Borghi, M., and Karapapa, S. 2011. Non-display uses of copyright works: Goggle books and beyond. *Queen Mary Journal of Intellectual Property* 1:21–52.

Börner, K.; Conlon, M.; Corson-Rikert, J.; and Ding, Y. 2012. *VIVO: A Semantic Approach to Scholarly Networking and Discovery*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

Del Rio, N.; Rosales, N. V.; Pennington, D.; Benedict, K.; Grady, C.; and Stewart, A. 2013. Elseweb meets sadi: Supporting data-to-model integration for biodiversity forecasting. In *Discovery Informatics Symposium*.

Gennari, J. H.; Musen, M. A.; Fergerson, R. W.; Grosso, W. E.; Crubzy, M.; Eriksson, H.; Noy, N. F.; and Tu, S. W. 2002. The evolution of protégé: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies* 58:89–123.

Gil, Y., and Hirsh, H. 2012. Discovery informatics: Ai opportunities in scientific discovery. In *2012 AAAI Fall Symposium Series*.

Grudin, J. 1994. Groupware and social dynamics: eight challenges for developers. *Commun. ACM* 37(1):92–105.

Maidment, D. 2013. Cuahsi's hydrologic information system (cuahsi-his). http://his.cuahsi.org/mastercvreg.html.

Moore, R. 2008. Towards a theory of digital preservation. *International Journal of Digital Curation* 3:63–75.

Porter, J. 2013. Lter controlled vocabulary. http://vocab.lternet.edu.

Salayandia, L.; Gates, A. Q.; and Pennington, D. 2013. Metashare: Constructing actionable data managmenet plans through formal semantics. In *Research Data Management Implementations Workshop*.

Salayandia, L.; Gates, A. Q.; and Pinheiro, P. 2012. An approach to evaluate scientist support in abstract workflows and provenance traces. In *Discovery Informatics Symposium: The Role of AI Research in Innovating Scientific Processes, AAAI Fall Symposium Series*.

Salayandia, L. 2012. *Ontologies for Scientific Data Transformation*. Ph.D. Dissertation, University of Texas at El Paso.

Simmhan, Y. L.; Plale, B.; and Gannon, D. 2005. A survey of data provenance in e-science. *SIGMOD Rec.* 34(3):31–36.

Wilkinson, M. D.; McCarthy, L.; Vandervalk, B.; Withers, D.; Kawas, E.; and Samadian, S. 2010. Sadi, share, and the in silico scientific method. *BMC bioinformatics* 11(Suppl 12):S7.