



A Case Study of Data Integration for Aquatic Resources Using Semantic Web Technologies

By Janice Gordon, Nina Chkhenkeli, David Govoni, Frances Lightsom, Andrea Ostroff, Peter Schweitzer, Phethala Thongsavanh, Dalia Varanka, and Stephan Zednik

Open-File Report 2015–1004

**U.S. Department of the Interior
U.S. Geological Survey**

U.S. Department of the Interior
SALLY JEWELL, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2015

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit <http://www.usgs.gov> or call 1–888–ASK–USGS

For an overview of USGS information products, including maps, imagery, and publications, visit <http://www.usgs.gov/pubprod>

To order this and other USGS information products, visit <http://store.usgs.gov>

Suggested citation:

Gordon, Janice, Chkhenkeli, Nina, Govoni, David, Lightsom, Frances, Ostroff, Andrea, Schweitzer, Peter, Thongsavanh, Phethala, Varanka, Dalia, and Zednik, Stephan, 2015, A case study of data integration for aquatic resources using semantic web technologies: U.S. Geological Survey Open-File Report 2015–1004, 55 p., <http://dx.doi.org/10.3133/ofr20151004>.

ISSN 2331-1258 (online)

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Although this information product, for the most part, is in the public domain, it also may contain copyrighted materials as noted in the text. Permission to reproduce copyrighted items must be secured from the copyright owner.

Contents

Abstract	1
Introduction.....	1
Challenges of Scientific Data Integration	1
The Potential of Semantic Web Technologies.....	2
Demonstration Project.....	3
Purpose and Scope.....	3
Use Case: Data Integration for Freshwater Fish Habitat Modeling.....	3
Data Sources.....	4
Methods.....	4
Application of the Methodology.....	6
Use Case Development and Iteration	6
Information Modeling.....	7
Technical Approach	9
Architecture Design Plan.....	9
Data Preparation	10
BioData.....	10
Multistate Aquatic Resource Information System	11
National Geochemical Survey.....	11
National Hydrography Data set.....	13
Prototype Development Approach.....	13
Results.....	14
User Interface Design	14
Prototype Architecture.....	15
TDB Triple-store	16
Fuseki SPARQL Endpoint.....	16
API Development.....	17
Integrated Data.....	18
Fish.....	20
Water.....	22
RDF Serialization in Turtle	23
Sediment.....	24
RDF Serialization in Turtle	24
Geospatial Data Integration.....	24
Discussion and Conclusions.....	25
Evaluation of the Prototype	25
Evaluation of the Use Case	26
Evaluation of the Methodology.....	27
Future Goals	27
Provenance and Data Quality.....	28
Technology Performance	28
Geospatial Semantic Integration	28

Use(s) of Linked Data.....	29
Acknowledgments.....	29
References Cited.....	30
Glossary.....	32
Appendix 1: Semantic Web Technologies (Overview).....	34
Data Interchange.....	34
Resource Description Framework.....	34
Resource Description Framework in Attributes (RDFa).....	34
Linked Data.....	34
Vocabularies.....	34
RDF Schema (RDFS).....	35
Simple Knowledge Organization Systems (SKOS).....	35
Web Ontology Language (OWL).....	35
Query.....	35
Rules.....	36
Rule Interchange Format.....	36
Semantic Web Rule Language.....	36
Appendix 2: Introduction to Resource Description Framework.....	36
Resource Description Framework.....	36
RDF Data Model.....	36
RDF Serializations.....	37
RDF/XML.....	38
N-Triples.....	38
Terse RDF Triple Language (Turtle).....	38
RDF Resources.....	38
Appendix 3: Linked Open Data.....	39
Open Data.....	39
Linked Data.....	39
Linked Open Data.....	39
Linked Open Data Resources.....	42
Appendix 4: API SPARQL Statements.....	43

Figures

1. Semantic technology development processes developed at Rensselaer Polytechnic Institute.....	5
2. An Observation as defined by the Observation and Measurements model.....	7
3. Observation describing the total length of a fish.....	8
4. Resource Description Framework serialization in Turtle syntax depicting an observation of the total length of a fish.....	9
5. High-level architecture developed during the technical-approach phase of the methodology.....	10
6. The user interface of the semantic data-integration prototype features dynamically updated widgets, which enable the user to interact with the system in an interactive dialog mode.....	14
7. Architecture diagram of the prototype system.....	16
8. This figure shows the namespace declarations of all the vocabularies referenced in the file containing the integrated data.....	19

9.	This figure represents an Observation for Sampling event PA76971, with the Feature of Interest being sample PA76971, the Procedure being a Trap Net, the Observed Property being Total Catch, and the Result being 1 fish	20
10.	This figure shows the RDF serialization in Turtle syntax of the data represented in figure 9, an Observation for Sampling event PA76971, with the Feature of Interest being sample PA76971, the Procedure being a Trap Net, the Observed Property being Total Catch, and the Result being 1 fish	21
11.	This figure represents an Observation for Sampling event PA1192, with the Feature of Interest being sample PA1192, the Procedure being a Secchi Disc, the Observed Property being Secchi Depth, and the Result being 0.89 meters	22
12.	This figure shows the RDF serialization in Turtle syntax of of the data represented in figure 11, an Observation for Sampling event PA1192, with the Feature of Interest being sample PA1192, the Procedure being a Secchi Disc, the Observed Property being Secchi Depth, and the Result being 0.89 meters.....	23
13.	This figure represents an Observation for Sampling event C-157189 for lead (Pb) Concentration, with the Feature of Interest being sample C-157189, the Procedure being ICP40, the Observed Property being Pb Concentration, and the Result being 19 parts per million (ppm).	24
14.	This figure shows the RDF serialization in Turtle syntax of the data represented in Figure 13, an Observation for Sampling event C-157189 for lead (Pb) Concentration, with the Feature of Interest being sample C-157189, the Procedure being ICP40, the Observed Property being Pb Concentration, and the Result being 19 parts per million (ppm).	24
15.	This figure shows the RDF serialization in Turtle syntax of the geoSPARQL query for a MARIS sampling site.....	25
2-1.	Graph depicting the triple where the subject is Fish, the predicate is Swims In and the Object is Lake	37
2-2.	Graph depicting the triple where the Subject is Lake, the predicate is Named and the object is Clear Lake	37
2-3.	A graph depicting two triples. The first triple has the Subject Fish, the predicate SwimsIn with the Object Lake. The second triple has the Subject Lake, the predicate is Named, and the Object Clear Lake	37
2-4.	Graphical representation of the triple where the subject is Fish, the predicate is SwimsIn, and the Object is Lake.....	38
3-1.	Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak	40
3-2.	A visualization of the steps needed to achieve a five-star data rating.....	42

Tables

1.	This table describes all of the available API parameters, definitions, and example values.....	17
2.	Example URL	18
3-1.	Description of Tim Berners-Lee's five star data model	41

A Case Study of Data Integration for Aquatic Resources Using Semantic Web Technologies

By Janice Gordon, Nina Chkhenkeli, David Govoni, Frances Lightsom, Andrea Ostroff, Peter Schweitzer, Phethala Thongsavanh, Dalia Varanka, and Stephan Zednik

Abstract

Use cases, information modeling, and linked data techniques are Semantic Web technologies used to develop a prototype system that integrates scientific observations from four independent USGS and cooperator data systems. The techniques were tested with a use case goal of creating a data set for use in exploring potential relationships among freshwater fish populations and environmental factors. The resulting prototype extracts data from the BioData Retrieval System, the Multistate Aquatic Resource Information System, the National Geochemical Survey, and the National Hydrography Dataset. A prototype user interface allows a scientist to select observations from these data systems and combine them into a single data set in RDF format that includes explicitly defined relationships and data definitions. The project was funded by the USGS Community for Data Integration and undertaken by the Community for Data Integration Semantic Web Working Group in order to demonstrate use of Semantic Web technologies by scientists. This allows scientists to simultaneously explore data that are available in multiple, disparate systems beyond those they traditionally have used.

Introduction

In the 21st century, earth scientists increasingly seek to integrate large data sets from multiple sources. This trend results from several factors: technical innovations that allow remote access and automated analysis of large volumes of data; the Nation's need for scientists to address problems that cross both the traditional boundaries of scientific disciplines and broader landscape scales; budget constraints that encourage use of the best available data rather than acquisition of new data to meet research requirements; and critical research problems that can be addressed only by combining data collected in the past to examine changes in earth systems, as well as the causes of those changes. The integration and scientific analysis of data from multiple sources can produce important insights, but also presents challenges. Members of the Semantic Web Working Group, within the USGS Community for Data Integration, have investigated the potential of Semantic Web technologies to address some of these challenges.

Challenges of Scientific Data Integration

Scientific knowledge resides with a community of people who use judgments accumulated and refined over years of experience. The scientific community increasingly relies on a network of computer systems to preserve and provide access to the scientific data that is the foundation for developing new knowledge. The scientific community is not a unified whole. Separate sub-communities specialize in

specific disciplines and localities. Each sub-community develops and uses its own techniques, concepts, and terminology best suited to its specific research problems. Sub-communities rely on systems to link customized databases with customized data-analysis tools. Widespread use of standards and service-oriented architectures often allows technology to integrate data relatively rapidly from multiple scientific domains. A scientist seeking to create a research data set by integrating data from multiple independent data systems frequently encounters obstacles, such as (1) trouble discovering the systems that have potentially useful data, (2) difficulty understanding and evaluating the data, and (3) complications in locating and extracting the relevant data elements along with the essential context and documentation.

Some of these obstacles arise because scientific sub-communities function as linguistic communities, each one expressing important distinctions through precise terminology that has been refined through its own internal scientific discourse, but perhaps never codified in a glossary. More difficult challenges arise because scientific use of the data requires preservation of domain-specific judgments and understandings that, within the sub-communities that created the data systems, might have been felt too obvious to mention. Standards for data interoperability frequently focus on a rudimentary approach that matches simple data elements but omits the annotations that are essential for scientific meaning. A repository of scientific observations may therefore use a Web service to provide standards-based access to data from analysis of environmental samples, but the service may fail to include the ancillary information that offers essential insight into the relevance of the data. For example, a sample might have been analyzed with a variety of methods, and the method used may determine the pertinence of the measured value for a given problem.

The Potential of Semantic Web Technologies

Semantic Web technologies offer the potential for addressing data-integration challenges that result from sub-community linguistic differences and unwritten assumptions. The basic idea of the Semantic Web is to encode meaning with the data so that automatic systems can operate appropriately without relying on human judgment. A simple text match would imply a relation of information about different entities that have similar names, but Semantic Web approaches encode additional information about what kind of entities they are. For example, a text search may combine in a single list Web sites that offer scientific data about the waters of Long Island Sound, those advertising hotels on the shore of Long Island Sound, and musicians offering to sing at weddings on Long Island. Semantic Web technologies are thus appropriate for clarifying the meaning of scientific terminology and documenting the assumptions that are built into domain-specific scientific databases. Possibly in the future a computerized expert system, containing a scientific knowledge base, could use Semantic Web technology to discover, evaluate, and integrate data into a research database. For now, Semantic Web offers the potential of (1) expressing the meaning of data so that a scientist can judge their suitability, (2) encoding documentation and context so that they are included in an integrated data set, and (3) providing a common data model Resource Description Framework (RDF) suitable for use with all data types.

In a way that is recognizable by automated processing, RDF resources identify anything that can be named: persons, places, things, concepts, relationships, and events. Most often, these are identified by unique internet addresses, which, in the best practice, provide information about the entity represented by the resource. Ontologies take the next step. Ontologies are networks of resources that enable machines to take actions consistent with human understandings of the relationships between things, for example using a resource that represents “is upstream of” to encode the relationships between

two sampling sites on a river. It is through these logical connections that the formal semantics of a resource are developed. (See Appendix 1 for more information on Semantic Web Technologies.)

Demonstration Project

This paper reports on a pilot project that developed semantic technologies to access multiple remote data systems, extract particular data values from each, and combine them into a single data set for download.

Purpose and Scope

The pilot project was sponsored by the Semantic Web Working Group of the USGS Community for Data Integration. The objectives of the Working Group are to demonstrate the use of Semantic Web technologies (1) to integrate multi-discipline data that were independently designed and created, and (2) to support the efficient use of information derived from the data by scientists whose investigations cross traditional scientific discipline boundaries and who are not data system specialists. Our work began with a demonstration project that had the specific purpose of developing a prototype to investigate the use of semantic technologies to relate and extract facts from several different data systems and to combine them into a new, conceptually consistent data set. This narrow focus is appropriate for the Working Group's demonstration project; it simplified the task by leaving several important aspects out of scope. We deferred the essential task of developing interfaces to link the new integrated data set with scientific-analysis applications. Ideally, the integrated data set would also include semantically enabled links to ancillary information associated with each of the original data sources, such as elaborations of data quality and provenance, but the creation of such linkages was outside the scope of this prototype.

The prototype addresses the technical challenge of enabling a scientist to find and download data. The prototype also addresses two key semantic aspects of scientific data integration. First, data values derived from observations must be documented with sufficient detail to enable scientists to assess the relevance, comparability, and quality of data from diverse sources. Second, the data must be reported and constrained using the geospatial and temporal relationships essential to the scientific study.

Technological obstacles to obtaining and using data can impede data integration; specific obstacles we have chosen to address include incompatibility within information models and inconsistency of semantic terms and data-processing parameters. A user interface that assists the user in navigating these incompatibilities from disparate data sources was deemed critical to the success of a data-integration system, and was included in the prototype design. Discussion of this online prototype product demonstrates the potential for Semantic Web technologies to identify comparable data from different sources, and illustrates how existing data models may be improved to facilitate greater interoperability.

Use Case: Data Integration for Freshwater Fish Habitat Modeling

The project was guided by a use case: data integration to address the requirements of freshwater fish habitat modeling. Effective prediction of the abundance of particular species at particular locations in a river or stream is a primary objective of scientific studies of general population dynamics and ecology. Managers of natural resource programs need better knowledge of fish ecology and aquatic habitat requirements, as well as improved tools for assessment and planning, to help conserve and rehabilitate populations throughout their native ranges. Within the U.S. Geological Survey (USGS), scientists working on the National Fish Habitat Action Plan (National Fish Habitat Partnership, 2012) and aquatic aspects of the USGS Gap Analysis Program (U.S. Geological Survey, 2012c) have these

goals: (1) develop empirical species-habitat models that effectively predict the potential of specific stream reaches as habitats for important fish species, (2) describe the predicted distribution of habitats of various qualities, and (3) compare predictions with observed fish abundances. The resulting models, data, and tools will help managers assess the condition of their stream-habitat resources and prioritize conservation efforts. Evaluation of the model structure and predicted habitat distribution will also provide insight into the suite of conditions that best support important fish species and how those conditions vary within and among watersheds. The research is currently conducted by discovering existing historical data and collecting new data, converting these data to compatible formats, and using GIS systems to combine and visualize the data and create a desired model. Semantic Web technologies have the potential to simplify the data-integration process and enable scientists to spend less time on data-integration procedures.

Data Sources

This project investigated semantic techniques to more effectively automate and expedite data discovery and integration from multiple sources and science disciplines, and thus to produce a robust information resource that project scientists can tap to create and evaluate their models. The project integrated data from multiple USGS data systems that were designed and created independently by different USGS offices to support a number of research goals, including those not specifically developed for fish habitat modeling:

- The **BioData Data Source** contains aquatic bioassessment data (biological community and physical habitat data) using the National Water Quality Assessment (NAWQA) protocol (U.S. Geological Survey, 2012a).
- The **Multistate Aquatic Resources Information System (MARIS)** is an on-line resource that provides single-point access to freshwater fish occurrence observations from multiple state fish and wildlife agencies (Multistate Aquatic Resources Information System, 2012).
- The **National Hydrography Dataset (NHD)** is the surface water component of *The National Map* of the USGS, and primarily supports mapping and flow modeling (U.S. Geological Survey, 2012e).
- The **National Geochemical Survey (NGS)** database provides data about the natural geochemical characteristics of the watershed as well the impacts of mining and refining operations along the stream and in its watershed (U.S. Geological Survey, 2012b).

Methods

We chose to use the Semantic Web Methodology and Technology Development Process created at Rensselaer Polytechnic Institute (RPI) (Fox and McGuinness, 2008).

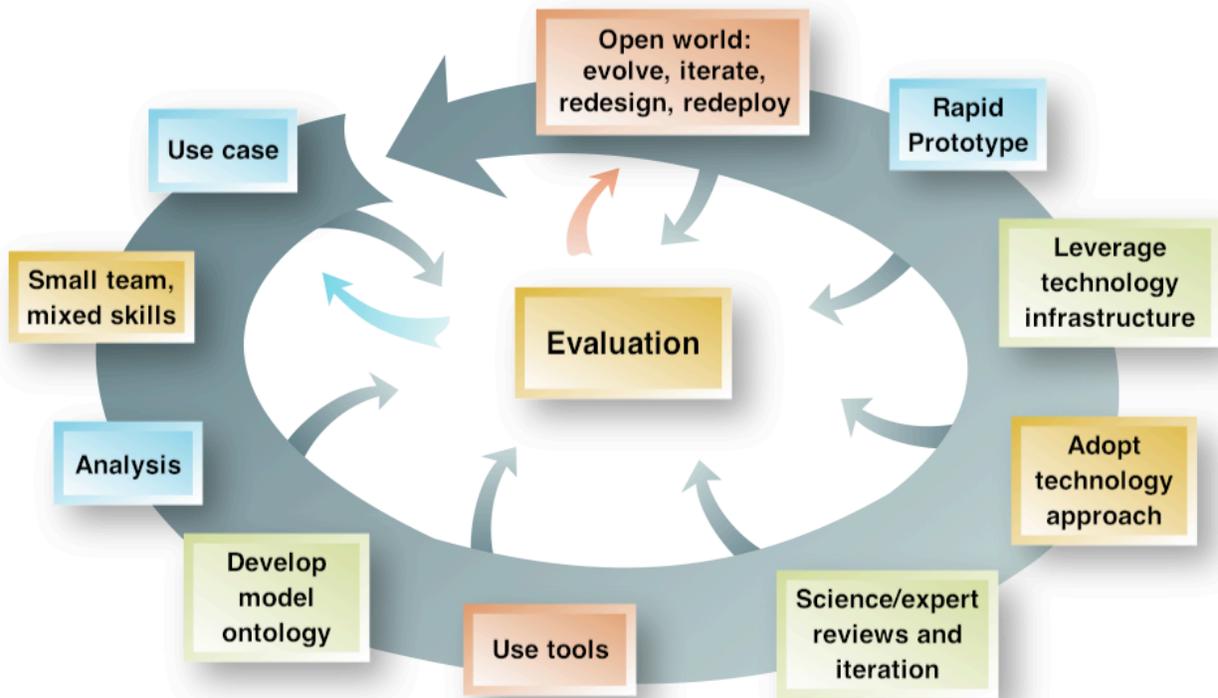


Figure 1. Semantic technology development processes developed at Rensselaer Polytechnic Institute (Fox and McGuinness, 2008).

The methodology is an iterative process, similar to a common software development life cycle, that starts with the development of a use case. A use case describes the interactions between a user and a system to achieve the user’s goal. The use case template for this methodology contains the following elements: goals, summary, actors, preconditions, triggers, basic and alternative flows, postconditions, activity diagram, and lists of the resources and services needed to build the system.

After one or more initial use cases are developed, a small team is formed with typically 5–12 members, each having a well-defined role. Team roles include: a facilitator, who has knowledge of the methodology and typical facilitation skills; domain experts, who have a knowledge of the resources, data, applications, and tools; technologists, who can help design the system and develop the code; data modelers, who extract concepts and relationships during information modeling; and the scribe, who has the important responsibility of writing everything down.

Once the team has been established, the members analyze each written use case. The analysis gives the team a greater understanding of the users and goals in the use cases. The analysis phase may also lead to a new iteration of the use-case documents after review and modification by the team, allowing the use case to be improved in areas that the team feels misrepresent the goals of a user. A use case is not expected to be “finished,” but rather brought to a point where it enables the team to proceed to the next phase in the development cycle.

When the use case is specified in sufficient detail, the team can then proceed to develop a model ontology. During this phase data modelers play a key role within the group by extracting important concepts from the use case, and defining the relationships between those concepts to create a conceptual data model. This conceptual model then becomes the basis for the model ontology. At this stage in the methodology, it’s important that the model is reviewed by domain experts and goes into an iterative

phase of revision and review before being implemented as a formal ontology within the software system.

The next phases in the methodology are to adopt a technology approach and find infrastructure that will support rapid prototyping of the software system. The team technologists use the system behavior described in the use cases to assess potential technological approaches, design a prototype architecture, and develop initial functional requirements. Rapid prototyping can then proceed using the initial functional requirements and the technical architecture. Once the first version of the prototype is complete, another evaluation and analysis phase begins. The use cases and early prototypes are used to further refine the functional requirements. The evaluation may then suggest additional use cases, which can be developed using the same methodology.

Application of the Methodology

Use Case Development and Iteration

The prototype design was driven by a data integration use case directed toward this overall goal: combine data from a variety of sources into a single, queryable data set to support aquatic habitat research of freshwater fish species in the Susquehanna River Basin. The summary of our use case is as follows:

Studies of aquatic fish ecology and habitat requirements depend on the availability and effective use of disseminated and heterogeneous data assets. Data often need to be integrated in order to build new knowledge about habitat conservation and rehabilitation. In this use case, a biologist needs to access and combine data about the Susquehanna River Basin that are currently contained in disparate databases. The data needed describe (1) the abundance of freshwater fish; (2) hydrology of the river basin region; (3) water quality and contaminant data; and (4) historical stream sediment geochemical data. The goal is to enable a modeler to understand the relation between fish populations and potential habitat contaminants in the Susquehanna River Basin through the combination of these data assets.

The use case team consulted a domain-expert scientist, who provided a user perspective on fisheries population occurrence data selection criteria. A key assumption defined through that consultation was that the user of the system possesses knowledge of the fish species and fish habitat pertinent to the geographic area for which the system will provide information. Data acquisition steps in the basic flow were predicated on this assumption. Data from fisheries, fish habitat, hydrographic, and stream sediment data sources were determined to be integral to the successful resolution of the use case scenario.

The development of the use case for our project was an iterative process that took place over the course of several months. We defined our primary actor as a fisheries biologist whose goal is to gather ancillary and baseline data to inform current or future research projects related to fish populations in the Susquehanna River Basin. We designed the basic flow to describe how the primary actor would interact with the system, choose and download data, and meet the goals defined in the use case. In addition to the basic flow, we identified several alternative flows, exception flows, and post conditions. Then we listed the resources required to build the system, which included data and services. Design of a proposed data-integration system to meet a particular use case such as this may help to discover new assumptions, but also results in a system that is biased to meet the needs of a particular group of users. To broaden the usefulness of our study, the use case description also included alternate flows involving a variety of perspectives, user objectives, and outcomes.

Our next step was to define an information model based on the concepts and relationships the user and system experience and use, as described in the use case basic flow. The information model

defines the domain concepts relevant to the scenario and user objectives. From our use case, it was clear that various spatial, temporal, observational, and process concepts would be part of our information model.

Information Modeling

Information modeling is a key step when integrating disparate data. The process of information modeling flows from conceptual to logical and then to physical data modeling as recommended by the American National Standards Institute (American National Standards Institute, 1975).

Conceptual modeling often begins with the creation of a concept map. The team used the Institute for Human and Machine Cognition (IHMC) Cmap Tools (Institute for Human and Machine Cognition, 2012) to design a diagram representing the basic ideas in the use case. The concept map was used to understand and document the key information concepts shared by the databases and relationships between them.

A master ontology was desired to allow the system to use a common information model to query and represent information from the differing data sources. The success of the data integration depended in part on aligning existing data source information with the derived common information model. Standardized vocabularies chosen to help guide the integration design include the International Standards Organization (ISO) Observation and Measurement (O&M) Model (Cox, 2010). The O&M model is based on the principles of observations and sampling. “An *Observation* is an action whose *result* is an estimate of the value of some *property* of the *feature-of-interest*, obtained using a specified *procedure*” (Solid Earth and Environment GRID, 2012). In this model, the term “observation” represents an event and is not to be confused with the data value, which is the result of the observation event. The feature of interest is the object of observation, or the feature that will be observed. The procedure is the process used to capture the data on the feature of interest. The observed property is the attribute being measured on the feature of interest. The result is the data that were captured during the measurement process (fig. 2).

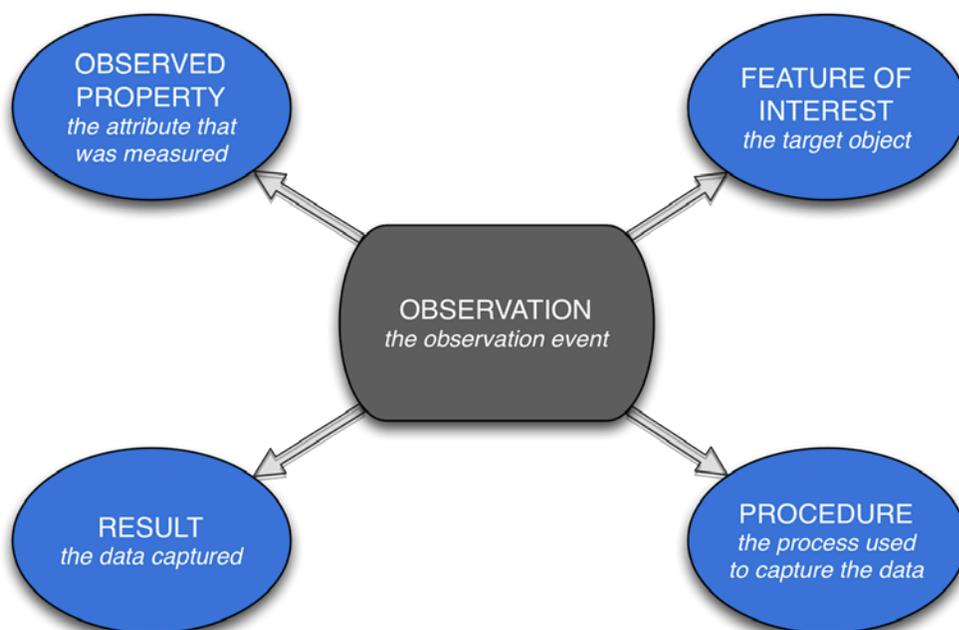


Figure 2. An Observation as defined by the Observation and Measurements model (Cox, 2010).

An example from one of our data sets applied to the O&M ontology model is the observation of the length of a fish. We can describe our example in three different ways, the first in a sentence, the second with a concept map, and third as RDF in Terse RDF Triple Language (Turtle) syntax (Beckett and others, 2012) using the O&M OWL ontology (Cox, 2011). Appendix 1 provides background information on RDF. In text, our example could be “the first fish specimen was caught using a minnow seine and the standard length was measured at 27 millimeters.” The same example as a concept map is figure 3, and in RDF Turtle shown in figure 4.

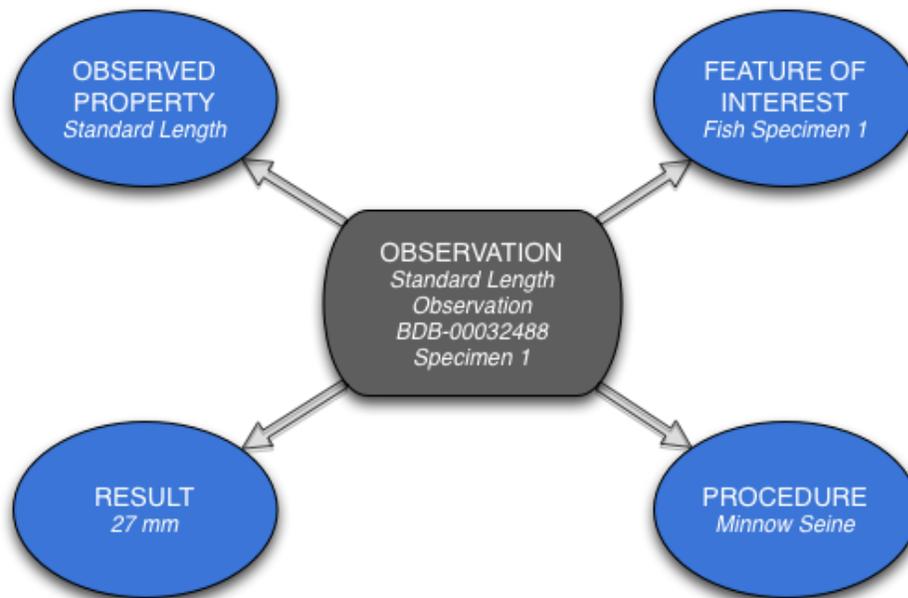


Figure 3. Observation describing the total length of a fish. (mm, millimeters)

```

<http://www1.usgs.gov/linkeddata/swwg/biodata/obs/
DB-000032488/specimen/1/total_length>
  a om:Observation, usgs:TotalLengthObservation;
  om:featureOfInterest <http://www1.usgs.gov/linkeddata/swwg/
biodata/sample/BDB-000032488/specimen/1>;
  om:observedProperty usgs:total_length ;
  om:result [
    a basic:Length, usgs:TotalLength;
    basic:number "27"^^xsd:float;
    basic:unit <http://www1.usgs.gov/linkeddata/swwg/unit/mm>
  ] .

<http://www1.usgs.gov/linkeddata/swwg/biodata/method/
BDB-000032488/MinnowSeine>
  a sam:Process.

<http://www1.usgs.gov/linkeddata/swwg/biodata/sample/
BDB-000032488/specimen/1>
  a sam:Specimen, usgs:Fish ;
  sam:samplingMethod <http://www1.usgs.gov/linkeddata/swwg/
biodata/method/BDB-000032488/MinnowSeine> ;
  usgs:fromCollection <http://www1.usgs.gov/linkeddata/swwg/
biodata/sample/BDB-000032488>;
  rdfs:label "fish #1".

```

Figure 4. Resource Description Framework serialization in Turtle syntax depicting an observation of the total length of a fish.

Technical Approach

The implementation team built the prototype system using three processes: (1) we designed a high level system architecture plan based on other semantic data-integration systems; (2) we proceeded with the preparation of the data for the prototype; and (3) we developed prototype code using a small development team.

Architecture Design Plan

Our high level system architecture plan included four basic elements: (1) a RDF database, called a triple store, for data storage, and (2) a web service, called a SPARQL endpoint, which allows applications to query information from the triple store using the SPARQL query language, (3) an application programming interface (API) is software that passes information from a web application to

the SPARQL endpoint to retrieve search results from the triple store, and (4) a client web application to allow end users to search for data through a user interface and download data through a web browser. (See fig. 5.)

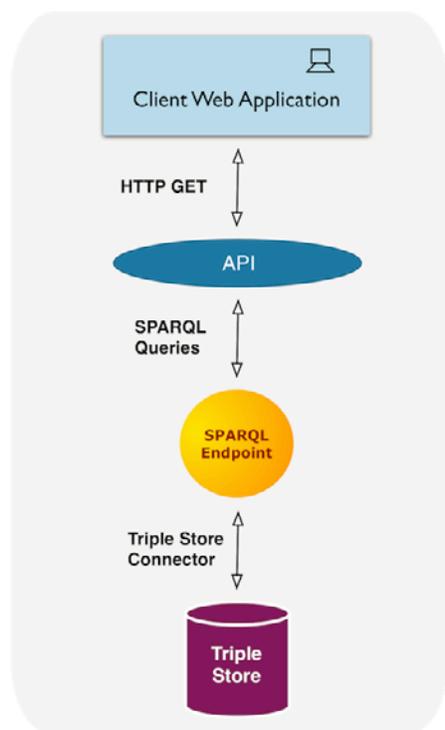


Figure 5. High-level architecture developed during the technical-approach phase of the methodology. Major elements include a Triple Store, SPARQL Endpoint, API, and a Client Web Application.

Data Preparation

After developing our high-level architecture, each data steward began the process of converting native relational data into RDF adhering to the O&M ontology. Descriptions of the processes used for converting each data set are below.

BioData

The USGS BioData Retrieval System (U.S. Geological Survey, 2012a) provides nationwide biological community and physical habitat data from stream ecosystems. The BioData database is stored in a relational/hierarchical model in its native format: Oracle RDBM tables. Each table in this hierarchical structure represents a domain object (for example Project, Fish Samples, Fish Counts, Study Reach, Fish Results), with each record containing instances of that domain object. At the top of the hierarchy are Projects with Fish Samples and Study Reaches as direct children. Existing further down the hierarchy are the children of Fish Samples; the Collection Methods, Fish Results, and Fish Counts. Additionally, Site Information is provided and linked to Fish Samples but is not explicitly a part of the hierarchy. Much of the challenge in converting this database to a triple store was identifying attributes from various levels of the data hierarchy to map into the O&M ontology. Related observations, observed properties, features of interest, procedures, and results (fig. 3) are stored across

different tables in the original data structure and had to be extracted and linked to form RDF that adequately expressed the same relationships.

The source data are provided in several formats, but for the purposes of transforming the data into RDF, the XML format was used for input. Python scripts were used to parse the XML and generate a set of RDF resources, with each resource getting a unique Uniform Resource Identifier (URI) assigned to it. These resources were then linked using both relationships defined in the O&M ontology and additional relationships designed to maintain the original relationships expressed in the source database. The Python scripts then serialized this new data structure into the Turtle format which was stored in a triple store.

Multistate Aquatic Resource Information System

The Multistate Aquatic Resource Information System (MARIS) data structure was developed by a consortium of State fish and wildlife agencies to allow states to provide a common set of variables, collected as part of aquatic species sampling surveys, and deliver them to stakeholders via the Web in a consistent format. The data available for fish-species occurrences, along with fish sampling methods and water quality measures at the time of collection, are geo-referenced with latitude/longitude coordinates. MARIS currently enables users to visualize summaries of the data available in tables and a browsable map. Data are available as Microsoft Access or Delimited Text files, though other formats are available as well.

The MARIS data were downloaded for New York, Maryland, and Pennsylvania in comma-delimited format. Each file was then transformed to RDF through TopBraid Composer's spreadsheet-import tool. The data from each state were then queried and limited to only the records found within the four-digit Hydrologic Unit Code for the Susquehanna River (0205). The last step in the data preparation was to translate the RDF to conform with the O&M ontology. Python scripts were written and run on each data file to generate RDF that complies with the O&M ontology as well as with the USGS elements identified in the data-modeling phase of the project. The O&M-compliant RDF was then loaded into another TDB triple store database.

National Geochemical Survey

The National Geochemical Survey (NGS) is a collection of stream sediment samples analyzed to determine the concentration of 42 chemical elements by strictly consistent laboratory methods. The database is intended to support regional mineral-resource assessments and exploration as well as environmental investigations of areas effected by mineral occurrences.

Samples were intended to represent the geochemical character of the areas from which they were drawn; materials eroded from a watershed collect in the stream sediments, so the chemical composition of those sediments represents the overall composition of the watershed. Sampling sites are distributed throughout the United States with a minimum spatial density of approximately one sample per 289 square kilometers.

Geographic location of the samples is generally well known. Many of the samples were collected originally during the late 1970s as part of the National Uranium Resource Evaluation (NURE) and were re-analyzed for the NGS to improve analytical precision. Since the NURE program was carried out prior to modern geospatial positioning systems (GPS), the locations are generally accurate but have variable precision. Samples newly collected for the NGS were located using GPS. The database containing these locations includes the county and state as well as the hydrologic unit (8-digit HUC) and 7.5-minute map quadrangle from which the sample was taken.

The NGS data are presented on the Web in a manner that emphasizes the sample, but we recognize that a potential user of these data may first ask whether a geochemical characteristic of particular interest was measured and by what analytical techniques it was assayed. Consequently, our approach included rearrangement of the data in a way that explicitly recognizes the individual chemical analysis as an independent item of interest to the scientist, and contains information about the sample-forming part of the detailed metadata surrounding that analysis.

The National Geochemical Survey database is stored in a PostgreSQL relational database and is available to the public for mapping, browsing, and download as part of the USGS Mineral Resources On-Line Spatial Data (MRData) Web site. The site provides interfaces by which subsets of a variety of geological, geochemical, and geophysical databases may be explored and downloaded for offline analysis. The display and download interfaces use the server-side PHP scripting language to draw information from the PostgreSQL database and present those data to users through HTML with some Javascript. Within the relational database-management system, the NGS data reside in a schema containing several tables; the principal scientific content of this database resides within a single table, with other tables providing metadata or supporting specialized indexes that speed search and download activities.

In this Semantic Web experiment, the procedure was to extract the data relevant to the prototype scenario from the database, translate that data into RDF conforming to the O&M ontology, and load the data into an RDF triple store used by the prototype system. Because MRData already has capabilities for selecting and accessing these data by geographic area, the plan was to create a module for re-expressing the data as RDF triples. This module was written in PHP and its export format is RDF-Turtle, a compact textual description of the relevant facts contained in the database. As noted above, this project's goals argued for a presentation of the geochemical information focusing on the observations themselves; that is, the geochemical analyses will be presented, rather than the collection of observations made on a given sample. In that way this project differs from the data-access philosophy employed by the host system MRData.

In addition to the observations and sample characteristics that reside in the NGS database (U.S. Geological Survey, 2012d), several types of RDF resources were created using information drawn from the metadata documenting the NGS, such as the concept of the concentration in a sample of a particular chemical species (here that generally refers to chemical elements rather than compounds), units of measure, types of material sampled, and analytic methods used to determine the chemical species concentrations. RDF describing the observations could then be written using these RDF resources rather than using literal values such as text. This strategy makes it possible to link those resources to corresponding characteristics of other, similar databases. These RDF resources were written in text files formatted as RDF-Turtle, and were passed to the prototype developers for entry into the triple store.

The NGS data also relate closely to a few other databases housed on MRData. Because most of the samples analyzed by NGS were originally collected as part of the National Uranium Resource Evaluation Hydrologic and Stream Sediment Reconnaissance program (NURE-HSSR), the observations and chemical analyses on those samples that were made by the NURE program should correlate closely with the NGS data, even though the analytical methods used in the NURE program were less precise due to advances in analytical technology over the past 30 years. Likewise, the MRData system includes a topically indexed bibliographic catalog of scientific data sets, of which the NGS database (U.S. Geological Survey, 2012d) is a record, and information from this topical catalog serves to link the NGS data to other databases related by geography or other scientific subjects. Consequently, RDF-Turtle descriptions of the NURE-HSSR data and the topical catalog were created to support further exploration of these and other related data sources by an explorer of the Semantic Web.

National Hydrography Data set

The NHD has nationwide coverage of waterbodies and drainage networks, called flowlines, and is organized into regions defined by the geographically nested Watershed Boundary Data set (WBD). Each region is identified by a hydrologic unit code (HUC) (U.S. Geological Survey, 2012e). In addition to the providing GIS geometry classes for mapping and other visualization, NHD supports waterflow modeling through the use of reach codes assigned to flowlines and waterbodies, and includes other features of interest represented as points. Because the NHD contains spatial data, there is inherent coordinate information, but the data model uses a linear referencing system whereby point data such as scientific observations are linked or referenced along one-one hundredth of the stream course length represented by the reach code (Simley and Doumbouya, 2012; U.S. Geological Survey, 2012). Geographic coordinates are not displayed in the attribute tables.

The Susquehanna River feature class and attribute tables for NHD and the WBD; these tables were converted to RDF using a custom program designed to convert Personal GDB .mdb files downloaded from The National Map Viewer interface. In the geodatabases, each NHD feature is represented by a row in the database table. Each type of feature has a table associated with it, for example: NHDPoint and NHDFlowline. For each table, a template was created that describes how to convert the values in the database row into triples. The patterns can also specify functions to be run that transform the column value. For this project, for example, FDate is converted to ISO 8601, the standard XML Schema Definition (XSD) date format.

Following the data conversion, taxonomic classes were created in an ontology of the NHD database to support the use of the data (Mattli and Viers, 2012). Several key aspects of the NHD database are documented in various unrelated sources, such as metadata, a user's guide, the data model diagram, or previous standards. The conversion allowed the relation of important information to automatically connect once the RDF instance resources were aligned with the NHD ontology. The geometry of features was encoded using Well Known Text (WKT), a standard of the International Standards Organization (International Standards Organization, 2011). The NHD RDF instance data and ontology have been made available via a SPARQL endpoint.

The O&M ontology does not include geospatial measurement units, although it references other ISO standards, which may have a controlled vocabulary for units. The connection between O&M and NHD is that NHD provides description for features, which are the overall feature of interest of observations from BioData, MRdata, and MARIS. The NHD would not contain any O&M observations itself, but acts as a way to describe the Features (Streams, Watersheds) connected to other observations. Other relations between the NHD and other data were through the use of the Open Geospatial Consortium GeoSPARQL standard (Perry and Herring, 2012). The GeoSPARQL "within" property was used to relate sampling sites to features described in the NHD RDF and to encode the location of the sampling site.

Prototype Development Approach

The development of our prototype began with the creation of a small team of three developers and one system administrator. Our team split development tasks into three components: (1) backend-services for the RESTful API, focusing on query and data return; user interface (UI) development, (2) the search interface and mapping features; and (3) system configurations, including data endpoint configuration and performance tuning.

Results

User Interface Design

Development of the UI for the system was guided by the use case that the project team developed early in the prototype-development process. Since the use case describes all user interactions with the system, it reflects the task flow for the user. Understanding of the task flow lays the foundation for the UI design.

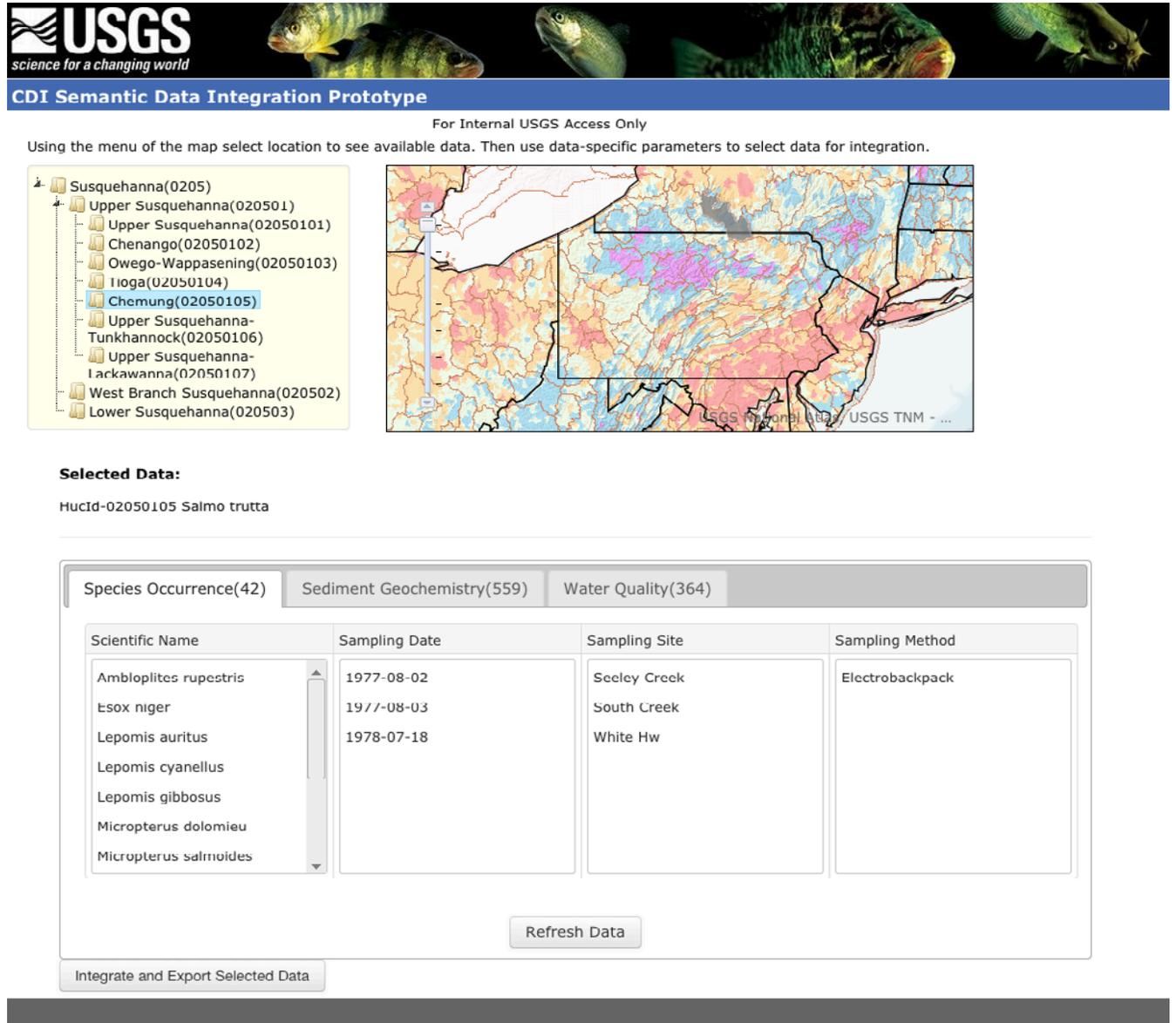


Figure 6. The user interface of the semantic data-integration prototype features dynamically updated widgets, which enable the user to interact with the system in an interactive dialog mode.

The use case for the demonstration project describes a multi-step process for the selection of data from disparate data sources for integration into a meaningful set of observation data with attributes,

which can be used for fish-habitat research. This kind of interaction with retrieval systems is best presented through a faceted search, which is a paradigm allowing users to explore a data source through interactions of refinement and expansion (Wagner and others, 2011). A faceted search is a multi-step filtering process. After the first interaction with the system, the user is offered a set of additional search parameters, or facets. These facets correspond to attributes of the information model. Facet contents describe related records or attribute values, so they help the user to assess what is available and also suggest potential additional filters. Every time the user adds or cancels a filter, the facets are updated to reflect the current set of results.

The concept of faceted search was developed outside of Semantic Web research, but has been widely adopted in interfaces of many semantic search systems. Faceted interface is effective in avoiding cognitive overload for users as they build complex queries. By gradual exposure of filters combined with constant system feedback, the interface guides the user in exploring available data and supports making informed selections at every step. This approach prevents “blind” choices, which could lead to null results.

Interaction with the prototype system starts with the selection of a location. At the stage of conceptual mapping, the team decided to normalize geospatial data from the heterogeneous data sources by mapping location information for all observation data to HUCs, representing watershed boundary areas that are used in the NHD. Since, in this classification, hydrologic units are geographically nested, the location selection in the interface is presented by a hierarchical tree menu that displays names of watersheds and their corresponding numeric codes. A map to the right of this menu allows the user to visualize the selected location. When a HUC is selected (highlighted) in the hierarchical menu, the map zooms on the selected unit and shows its boundaries outlined in bold.

The use-case objective was to make a data-integration system instead of a population occurrence data portal. We chose to develop a system that dynamically generates an integrated data product containing observational and sampling information selected by the user from disparate data sources. The interface supports this by displaying available observational data types in a tabbed content area, where each tabs presents a data type: Species Occurrence, Water Quality, and Sediment Geochemistry (fig. 6). After the location of interest is selected, the tab view is refreshed to show the number of observations of each data type available for this location. If there are no data for an observation type that is of interest to the user, another location can be selected and explored for data availability. The contents within each tab area are filled with “facet widgets.” These provide lists of values for attributes that characterize the available data and can serve as additional filters. For example, attributes for Species Occurrence data in the current version of the prototype are Scientific Name, Sampling Date, Sampling Site, and Sampling Method. Development of wireframes for the prototype required an inventory of all potential user choices, but only attributes critical for assessing a data set for integration were communicated through the interface. All selected query parameters are shown in the query status bar, which is placed in the center of the screen. As the user refines the query, the status bar is dynamically updated. Upon completion of the exploration and selection process, the user can export all data as an integrated RDF file.

Prototype Architecture

We chose to use the open source Apache Jena framework for our prototype (Apache Jena, 2012a). Jena is a Java framework for building Semantic Web applications. The framework provides extensive Java libraries that handle RDF, RDFS, RDFa, and OWL, and execute SPARQL queries over RDF. The specific Jena components used in this project are outlined below.

TDB Triple-store

A triple-store is a type of database used for the storage and retrieval of RDF triples. The TDB triple-store (Apache Jena, 2012b) was used to store the RDF created from BioData, MARIS, and the National Geochemical survey. The National Hydrography data set RDF was referenced by the prototype system from an external SPARQL endpoint hosted by The National Map program.

Fuseki SPARQL Endpoint

Fuseki is a SPARQL server that provides REST-style SPARQL updates and queries over HTTP. Fuseki supports (Apache Jena, 2012c) the following World Wide Web Consortium (W3C) standards: SPARQL 1.1 QUERY, SPARQL 1.1 Update, SPARQL 1.1 Protocol, and the SPARQL 1.1 Graph Store HTTP Protocol.

We created three Fuseki endpoints, one for each local TDB triple-store. Each public endpoint can process queries from the prototype application, as well as external applications thus providing a Linked Open Data source. Each endpoint can process SPARQL queries and return RDF results as XML, JavaScript Object Notation (JSON), or text format.

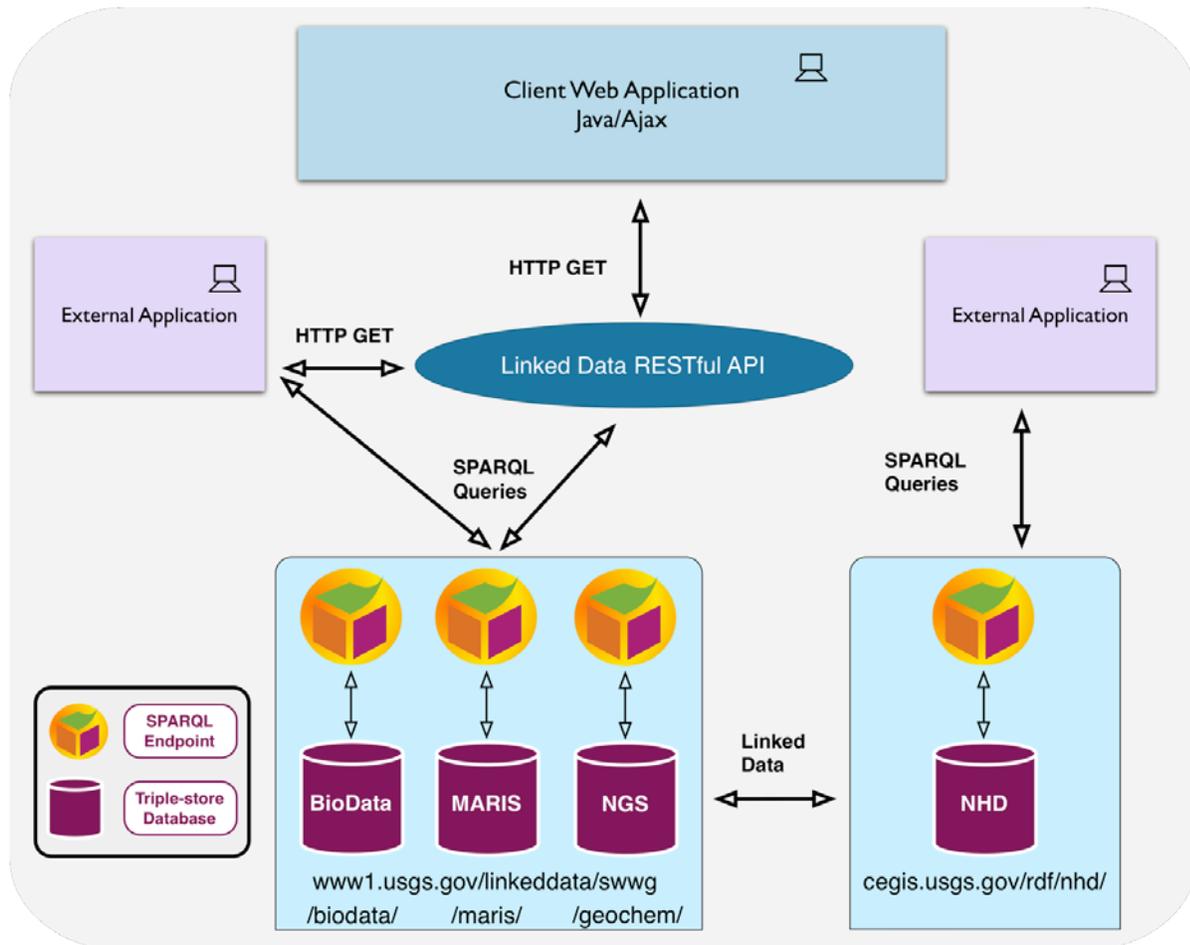


Figure 7. Architecture diagram of the prototype system. (API, application programming interface; BioData, Aquatic Bioassessment Data for the Nation; MARIS, Multistate Aquatic Resources Information System; NGS, National Geochemical Survey; NHD, National Hydrography Dataset)

API Development

An API is used to provide programmatic access to data or services provided by another system. The API describes the rules of the interaction so that a programmer can write code to extract data from the system in a prescribed and standardized manner.

Our original architecture included the use of Epimorphics implementation of the Linked Data API (ELDA) (Epimorphics, 2012) as a framework for implementing our desired API; however, during development of the back-end services we encountered difficulties using ELDA with desired service calls, so we chose to implement our API using custom Java code. We developed a RESTful service that processes query parameters in the URL and translates them into SPARQL queries to execute against the various data-source endpoints. Table 1 shows all of the available parameters as well as the definitions and examples defined in our API to query data from each local triple-store. The API design and parameter selection was driven by the use case and the search-interface design requirements outlined for the prototype application.

Table 1. This table describes all of the available API parameters, definitions, and example values. (ITIS, Integrated Taxonomic Information System)

Parameter	Definition	Example
huc	hydrologic unit code	02050104
so_tsn	species observation ITIS taxonomic serial number	162003
so_sampling_date	species observation sampling date	2005-10-14
so_sampling_site	species observation sampling site	Beechwood Lake
so_sampling_method	species observation sampling method	Electrobackpack
wq_water_characteristic	water quality observation characteristic	ph
wq_sampling_date	water quality characteristic sampling date	1998-08-15
wq_sampling_site	water quality sampling site sediment	Bear Creek
sg_chemical_species	geochemistry chemical species sediment	Se
sg_sampling_site	geochemistry sampling site sediment	PATJ143S1
sg_sampling_date	geochemistry sampling date	1977-09-15

An example of a potential user need can help to illustrate how a query is executed. A user is interested in finding information on chain pickerel in Beechwood Lake, Pennsylvania, and also wants any ancillary information about the water quality in the lake and regional sediment chemistry, specifically pertaining to selenium concentrations. The user would choose selections within the prototype's faceted search interface, which results in the following URL: `http://<baseurl>/occurrence_data?huc=02050104&so_tsn=162143&so_sampling_site=Beechwood%20Lake&sg_chemical_species=conc:Pb&wq_sampling_site=Beechwood%20Lake`.

The URL query parameters are explained in the table below.

Table 2. Example URL

(http://<baseurl>/occurrence_data?huc=02050104&so_tsn=162143&so_sampling_site=Beechwood%20Lake&sg_chemical_species=conc:Pb&wq_sampling_site=Beechwood%20Lake) broken down by parameter, value and a more detailed explanation.

Parameter	Value	Explanation
huc	02050104	the hydrologic unit code for the Tioga
so_tsn	162143	The observed species' taxonomic serial number from the Integrated Taxonomic Information System TSN 162143 represents the species <i>esox niger</i> , commonly known as the chain pickerel
so_sampling_site	Beechwood Lake	the sampling site of the species observation
sg_chemical_species	conc:Pb	concentration of lead (Pb) in the sediment samples
wq_sampling_site	Beechwood Lake	available water sample characteristics measured at the Beechwood Lake sampling site

Now that the user has specified the search parameters of interest, the API constructs and executes a SPARQL query based on the API call and the query parameters. Each SELECT statement uses the query parameters that were specified in the URL as SPARQL FILTERS when querying each endpoint. The RDF data returned from both CONSTRUCT statements is then concatenated through custom Java code and made available for download as a single RDF file serialized as RDF/XML. Appendix 4 shows a close examination of each of the SPARQL statements produced from the query parameters.

Integrated Data

The user is provided the option to download an RDF file that contains an integration of content from the disparate data sources and is compiled by aggregating the results of construct SPARQL queries run on the different data sources and filtered by current user selections. The integrated data file contains data from the original data sources, but now they are expressed using the Observations & Measurements information model (Observations & Measurements Ontology) and combined using the RDF data model.

Below is a review of a several code snippets from the RDF returned in the previous example. First is an examination of the RDF file namespace declarations referencing the different vocabularies utilized.

```
@prefix basic: <http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#> .
@prefix conc: <http://mrdata.usgs.gov/geochem/concentration#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix geo: <http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/> .
@prefix method: <http://mrdata.usgs.gov/geochem/method#> .
@prefix ngs: <http://mrdata.usgs.gov/geochem/field#> .
@prefix nure-site: <http://mrdata.usgs.gov/nuresed/site/> .
@prefix om: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sam: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#> .
@prefix usgs: <http://www1.usgs.gov/linkedata/usgs#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

Figure 8. This figure shows the namespace declarations of all the vocabularies referenced in the file containing the integrated data.

This file combines commonly used vocabularies such as Dublin Core with custom vocabularies such as the NGS and NURE. The code snippet in figure 8 shows the namespace declarations for each vocabulary. For example, Dublin Core is using the dc namespace, NGS is using the ngs namespace, and NURE is using the nure-site as its namespace (Dublin Core, 2013). This highlights RDF's flexibility, enabling vocabularies to be mixed and matched as needed when integrating disparate types of data. Different data sets may have points of integration along geospatial or temporal elements, but differ dramatically among elements such as collection methodologies. This prototype demonstrates the ability of Semantic Web technologies to combine different types of data (such as data pertaining to water quality, species occurrence, and sediment geochemistry) into a common model while still allowing the individual differences among the data to be expressed. This example shows that water-quality characteristics, chemical concentrations, and species occurrence sample data have all been described using the O&M concept of an observation. Each type of data can be combined using this common language, even though the features observed are drastically different. A closer look at the idea of an observation illustrates how this can apply to fish, water, and sediment samples.

Fish

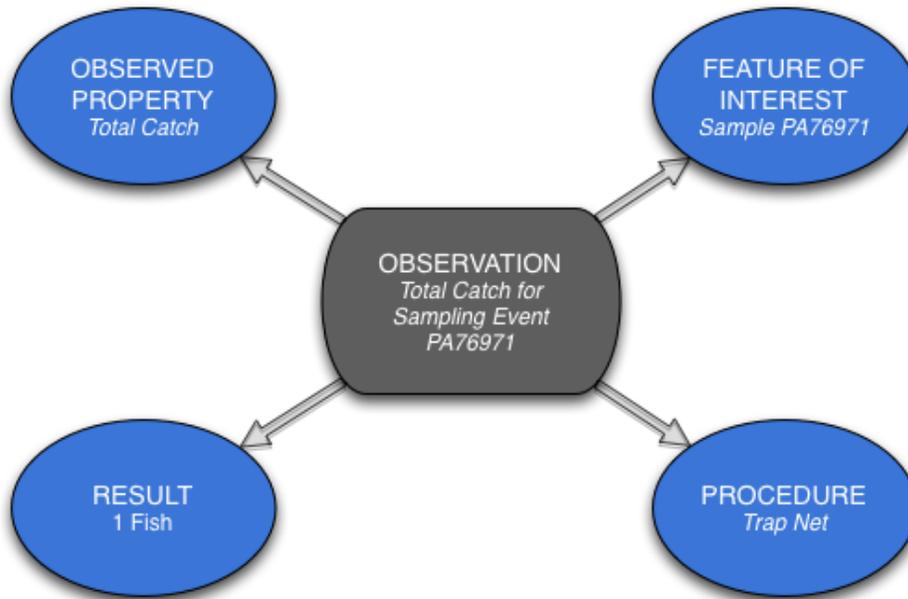


Figure 9. This figure represents an Observation for Sampling event PA76971, with the Feature of Interest being sample PA76971, the Procedure being a Trap Net, the Observed Property being Total Catch, and the Result being 1 fish.

```

<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/
PAI415117773045/sample/fish_PA76971>
  sam:relatedObservation
    <http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/
PAI415117773045/sample/fish_PA76971/obs/total_catch> ;
  sam:samplingLocation
    <http://www1.usgs.gov/linkedata/swwg/maris/site/pasite_415118E11> ;
  sam:samplingMethod <http://www1.usgs.gov/linkedata/swwg/maris/
method/PAI415117773045/TN> ;
  sam:samplingTime _:b3 ;
  usgs:effortTime _:b4 ;
  usgs:itisTsn <http://www1.usgs.gov/linkedata/swwg/tsn/162143> ;
  usgs:marisId "PAI415117773045"^^xsd:string ;
  usgs:stateSpeciesId <http://www1.usgs.gov/linkedata/swwg/state/
pennsylvania/species/195> ;
  usgs:targetStandard "ALL"^^xsd:string ;
  usgs:waterId "415118E11"^^xsd:string .

<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/
PAI415117773045/sample/fish_PA76971/obs/total_catch>
  om:observedProperty <http://www1.usgs.gov/linkedata/swwg/property/
total_catch> ;
  om:result
    [ basic:number 1 ;
      basic:unit <http://www1.usgs.gov/linkedata/swwg/units/number-of-fish>
    ] .

```

Figure 10. This figure shows the RDF serialization in Turtle syntax of the data represented in figure 9, an Observation for Sampling event PA76971, with the Feature of Interest being sample PA76971, the Procedure being a Trap Net, the Observed Property being Total Catch, and the Result being 1 fish.

Water

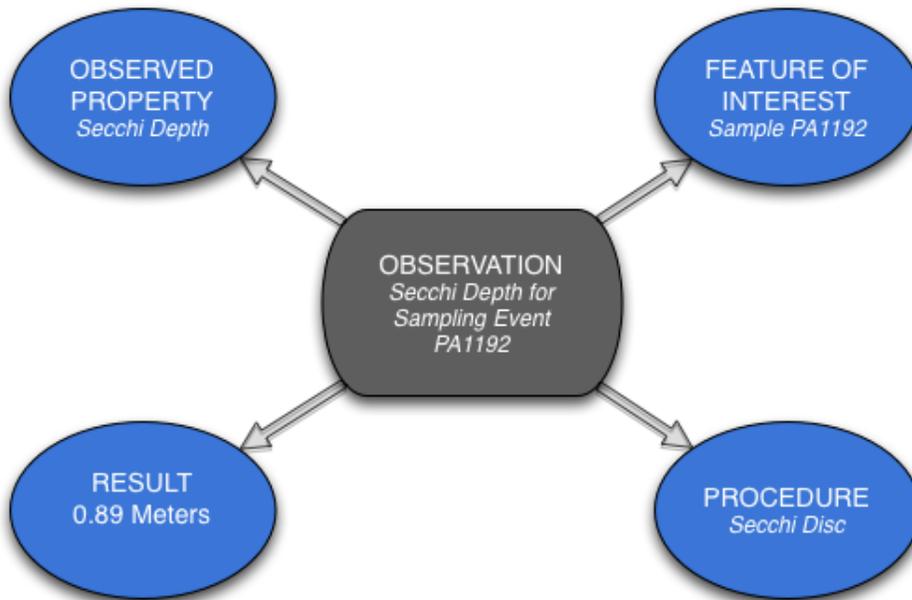


Figure 11. This figure represents an Observation for Sampling event PA1192, with the Feature of Interest being sample PA1192, the Procedure being a Secchi Disc, the Observed Property being Secchi Depth, and the Result being 0.89 meters.

RDF Serialization in Turtle

```
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/
sample/water_PA1192>
  sam:relatedObservation
    <http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/
PAI415117773045/sample/water_PA1192/obs/alkalinity> , <http://www1.usgs.gov/
linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1192/
obs/ph> , <http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/
PAI415117773045/sample/water_PA1192/obs/water_temp> , <http://www1.usgs.gov/
linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1192/
obs/secchi_depth> , <http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/
PAI415117773045/sample/water_PA1192/obs/conductivity> ;
  sam:samplingLocation
    <http://www1.usgs.gov/linkedata/swwg/maris/site/pasite_415118E11> ;
  sam:samplingTime
    [ usgs:begin "1988-08-16"^^xsd:date ;
      usgs:end "1988-08-16"^^xsd:date
    ] ;
  usgs:marisId "PAI415117773045"^^xsd:string ;
  usgs:waterId "415118E11"^^xsd:string .

<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/
sample/water_PA1192/obs/secchi_depth>
  om:observedProperty <http://www1.usgs.gov/linkedata/swwg/property/
secchi_depth> ;
  om:result
    [ basic:number "0.89"^^xsd:integer ;
      basic:unit <http://www1.usgs.gov/linkedata/swwg/untis/meters>
    ] .
```

Figure 12. This figure shows the RDF serialization in Turtle syntax of the data represented in figure 11, an Observation for Sampling event PA1192, with the Feature of Interest being sample PA1192, the Procedure being a Secchi Disc, the Observed Property being Secchi Depth, and the Result being 0.89 meters.

Sediment

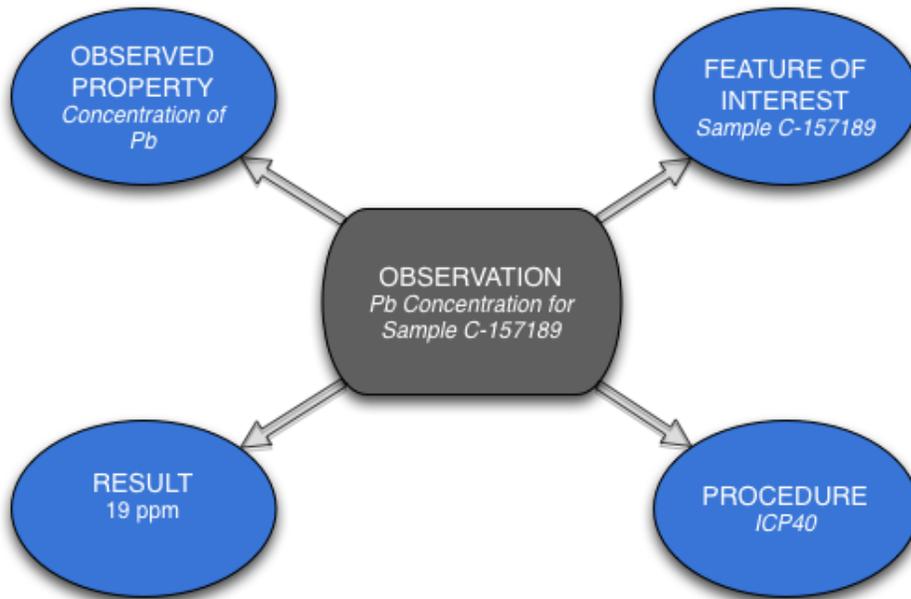


Figure 13. This figure represents an Observation for Sampling event C-157189 for lead (Pb) Concentration, with the Feature of Interest being sample C-157189, the Procedure being ICP40, the Observed Property being Pb Concentration, and the Result being 19 parts per million (ppm).

RDF Serialization in Turtle

```
<http://mrdata.usgs.gov/geochem/observation/C-157189/pb_icp40> a om:Observation;  
om:featureOfInterest <http://mrdata.usgs.gov/geochem/sample/C-157189>;  
om:observedProperty conc:Pb;  
om:procedure method:ICP40;  
om:result [  
  basic:number 19;  
  basic:unit units:ppm ];  
dcterms:isPartOf <http://mrdata.usgs.gov/catalog/catalog/29> .
```

Figure 14. This figure shows the RDF serialization in Turtle syntax of the data represented in figure 13, an Observation for Sampling event C-157189 for lead (Pb) Concentration, with the Feature of Interest being sample C-157189, the Procedure being ICP40, the Observed Property being Pb Concentration, and the Result being 19 parts per million (ppm).

Geospatial Data Integration

This project used spatial information as a dimension of integration. All observations are associated with a location and all observation locations have been associated with an NHD HUC. GeoSPARQL vocabulary “within” property is used to encode the location of the sampling sites to HU

features described in the NHD RDF. GeoSPARQL vocabulary tests the topological relations of two geospatial feature geometries based on the binary intersections of their interior, boundary, or exterior regions. The “geo:within” relation occurs when one feature geometry shares the interior space with another, but not its boundary or exterior. See figure 15 for an example of how MARIS uses GeoSPARQL.

```

<http://www1.usgs.gov/linkedata/swwg/maris/site/pasite_412402E11> a gf:feature,
usgs:StreamSamplingSite;
rdfs:label "Trout Rn" ;
usgs:waterName "Trout Rn"^^xsd:string ;
usgs:waterID "Trout Rn"^^xsd:string ;
usgs:waterType "STREAM"^^xsd:string ;
usgs:collectionAccuracyDescription "UNKNOWN"^^xsd:string ;
usgs:longitude [ a basic:Measure, usgs:Longitude;
basic:number "-77.54"^^xsd:float;
basic:unit <http://www1.usgs.gov/linkedata/swwg/unit/dd> ];
usgs:latitude [ a basic:Measure, usgs:Latitude;
basic:number "41.42"^^xsd:float;
basic:unit <http://www1.usgs.gov/linkedata/swwg/unit/dd> ];
geo:within <http://cegis.usgs.gov/rdf/nhd/hucf/02050205> ,
<http://www1.usgs.gov/linkedata/swwg/state/pennsylvania> ;
usgs:marisId "PAS412401772743414277543333"^^xsd:string .

```

Figure 15. This figure shows the RDF serialization in Turtle syntax of the geoSPARQL query for a MARIS sampling site.

The property geo:within is used to connect this StreamSamplingSite to the NHD resource. This site location is specified as a geo:Point and the site is specified to be within the specific HU 0205020 and the state of Pennsylvania.

Discussion and Conclusions

The final stage of the RPI methodology (fig. 1) is “Open world: evolve, iterate, redesign, and redeploy.” At this stage we evaluate the prototype in meeting the goals of the use case. The stage also includes identifying changes in the use case itself to better meet the goals of the project in a new development cycle. In a larger context, we also consider whether use of the RPI methodology met the goals of SWWG in sponsoring the project.

Evaluation of the Prototype

The prototype met the use-case goal: to combine data from a variety of sources into a single data set to support aquatic habitat research of freshwater fish species in the Susquehanna River Basin. The faceted interface is effective in allowing users to investigate the data available from the three systems and choose a location, time, and set of observations that produces a coherent data set of a reasonable size. The chosen data values are extracted and combined in a single RDF file that contains data and

metadata from the original sources in a uniform RDF information model (Observations & Measurements Ontology). The prototype demonstrated the ability of Semantic Web technologies to combine different types of data into a common model while still allowing the individual differences among the data to be expressed.

The development of the prototype provided an opportunity for team members to learn about many of the core Semantic Web tools and technologies and their application for USGS needs. In this regard, the prototype development fulfilled the SWWG goal of learning Semantic Web technologies and how they can support the USGS mission. It gave the team insight into the challenges of building a functional tool that will integrate data from diverse existing USGS data systems.

We encountered our first challenge when converting native data formats, such as relational and spreadsheet data, into RDF format that was consistent with the O&M Ontology. Most team members were unfamiliar with both the O&M ontology and the process for converting data into RDF. Several iterations were made for each data set needed for the prototype. The accuracy of the conversion methods and the resulting data still need further quality control and testing before data will be suitable for public release.

Our second major challenge came when attempting to utilize the Epimorphics implementation of the Linked Data API (ELDA) for our prototype system. We originally thought that this API implementation would help simplify our development; however, we soon realized that ELDA wasn't going to work for querying multiple SPARQL endpoints simultaneously through a single API configuration. The ELDA API implementation was designed to function as a simple, customizable API for interactions with a single triple store. There might be a way to modify and extend ELDA's functionality for querying multiple triple stores. However, little documentation was found on the subject, so the team proceeded with creating a custom API.

A third major challenge was slow performance processing SPARQL queries at several of the Fuseki endpoints. There may be several factors contributing to the performance issues, such as configuration issues and improper system tuning of the endpoints and triple-stores, poor data modeling and improper RDF implementation, inefficient SPARQL queries, or a combination of all these factors. The team wasn't able to investigate these issues due to the short time frame for completing the project. The interim solution was to limit the size of the data search to 100 records from the user interface.

Despite all of these challenges, the team was able to successfully achieve the major objectives outlined in our use case with the completion of the prototype application. A user is able to define a hydrologic unit code, choose observation data types of interest, and download a valid RDF/XML file containing data from each relevant data source.

Evaluation of the Use Case

The intent of the project was to demonstrate the use of Semantic Web technologies to select data from four independent USGS data systems and combine them into a single data set in RDF format. The use case was successful in focusing the team's attention on incompatibility between the systems' information models and inconsistency of semantic terms and processing parameters. For this first iteration, the use case was simplified by leaving out some important elements that should be included in the next cycle. One element has already been mentioned: the quality requirement that the interface return search results reasonably quickly without an artificial limit on the number of records. A second missing element is true integration of the temporal and geospatial relationships important to the stream habitat research: upstream/downstream, previous/subsequent in time, and collocation within a stream reach. The next version of the use case should make more use of the information available in NHD to meet these scientific requirements. Finally, the next use case should be more explicit about the

provenance and context information which is needed for scientists to have confidence in combining and comparing data from diverse sources.

Evaluation of the Methodology

In sponsoring the project, the SWWG hoped to learn Semantic Web techniques and demonstrate their potential to (1) express the meaning of data so that a scientist can judge the data's suitability, (2) encode documentation and context so that it is included in an integrated data set, and (3) provide a common data format (RDF) suitable for all data types. The project made progress on all three goals. The prototype produces an integrated data set in RDF. The user interface employs a faceted approach that allows a scientist to investigate the data available from three different systems, but at this point fails to provide previews or other means to evaluate data suitability. To make quick progress on the prototype, the team demonstrated that it is possible to include metadata parameters in the integrated data set, but only a few easily available parameters rather than fully integrated documentation and context that would support the scientific use of the data.

The semantic-technology development methodology starts with the development of a use case. At this stage of the approach, intensive work to clarify a sound plan was required to meet our objectives. The goal of our team was to test semantic technologies using predefined data sets, so our difficulty in defining a use case primarily stemmed from the fact that we were looking for a problem that could be solved by our data instead of looking for a solution to an existing science question. This goes against the true purpose of use case development and slowed down the development process.

Once the team overcame the challenges of defining a use case, the methodology provided a sound road map to guide us through our Semantic Web technology investigation. The information-modeling phase gave the team a greater understanding of each data set's structure, the similarities and differences among the data, and allowed us to define points of integration that would be used in our prototype. The methodology focused our efforts on the technical challenges of the technology approach, infrastructure leveraging, and prototyping phases (refer to Semantic Web Methodology and Technology Development Process diagram in fig. 1), which were occasions for meeting the Semantic Web Working Group goal of learning to build Semantic Web technologies.

The use-case methodology was especially helpful in designing the user interface. Since the use case describes all user interactions with the system, it reflects the task flow for the user, and understanding of the task flow lays the foundation for the user-interface design.

Future Goals

The objectives of the Semantic Web Working Group are to demonstrate the use of Semantic Web technologies by (1) integrating multi-discipline data that were independently designed and created and (2) supporting the efficient use of information derived from the data by scientists whose investigations cross traditional scientific discipline boundaries and who are not data-system specialists. Each step along the way throughout this investigation of Semantic Web technologies and development methodology has given the team an opportunity to learn the value and the challenges of real-world implementation to address science questions using USGS data. Our investigation answered some questions, but opened the door for even more work in the future. We identified four goals for future investigation: 1. Address Provenance and Data Quality; 2. Improve System Performance; 3. Explore Deeper Geospatial Semantic Integration; and 4. Expand the Use(s) of Linked Data.

Provenance and Data Quality

Issues of Provenance and Data Quality were deemed out of scope for our prototype project; however, we acknowledge the importance of providing this information with the resulting integrated data set. In future iterations of our prototype, we will need to address how provenance and data quality can be described for data modeled and formatted using Semantic Web technologies and standards. In order to address these concerns, the team would like to explore the use of the [Open Provenance Model \(OPM\)](#), [Open Provenance Model Vocabulary \(OPMV\)](#), or the [PROV-O ontology](#) for documenting the origin and transformation of the data used in our prototype. The Open Provenance Model allows for the exchange of provenance data between systems. The vision for the model states that it “allows provenance from individual systems to be expressed, connected in a coherent fashion, and queried seamlessly” (Moreau and others, 2010).

The Open Provenance Model Vocabulary (OPMV) is based on the Open Provenance Model and provides a lightweight vocabulary of terms that can be combined with vocabularies such as Dublin Core to help data owners to publish data that includes provenance information.

In order for our prototype system to become a successful production level data-integration system, the provenance of the data must be included. Responsible data-management practices are critical to enable the use and reuse of data in earth science research and our future work must address the techniques for describing provenance data within semantic data systems.

Technology Performance

The development of the prototype revealed significant performance issues that the team was unable to address during the first phase of development. The technical team needs to investigate the slow query responses found in the prototype system. Several possible areas in need of optimization include: the SPARQL queries, the configuration of the TDB triple-store, the configuration of the SPARQL Endpoint, and the quality and consistency of each data source. Each of these areas could be contributing to the slow performance of the system and needs to be investigated by the technical team. Additional or alternative technologies, such as different types of triple-stores, D2R servers, and other SPARQL endpoint implementations, could also be evaluated as potential solutions to our performance problems.

Geospatial Semantic Integration

Geospatial semantics were limited to HUC 8; this limitation omitted more granular levels where specific sampling sites are best represented. The NHD was used in the prototype primarily to reference features such as sites and watersheds together, but sampling site data were not encoded on maps. Such mapping could be done in the future by applying the WKT geometry serializations of features from NHD together with the geometric location references for other data. A benefit provided by referencing the NHD URI in the prototype is that it is ready for new linked data and ontology restrictions.

Expanded use of attribute data could produce variables of interest to the users, such as streamflow modeling. The NHD can be used to compute the size of the stream reach and its distance upstream from the mouth of the stream, and it also provides information about headwaters and tributaries. An intermediary ontology to integrate the NHD data model and user needs, similar to the role the O&M ontology served for observation data, would be needed. The core of the O&M ontology, called Observation Core, is deliberately left without the specification of location, as observations of some types are either devoid of location or are studied with only proximate location. Integrating specific spatial data, such as the NHD, with the O&M ontology involves the Specialized Observation module of

the standard. The Specialized Observation module depends on ISO 19107 Spatial Schema for geometry and topology. This initial pilot project of the Community for Data Integration (CDI)-funded Semantic Web Working Group did not involve extensive spatial data integration to map, for example, location or represent context for the field observations, but these spatial aspects will be part of the next stage of the study. These essential relationships include upstream/downstream and collocation within a stream reach. Another challenge for integrating the data for science research lies in recording the temporal aspects of the observations. On this point, the O&M standard depends upon “[ISO 19108, Temporal schema](#),” and “[ISO 19123, Schema for coverage geometry and functions](#).”

Use(s) of Linked Data

The team realizes the importance of expanding the use of linked data—specifically linked open data—for USGS data sources and data-analysis systems. Linked data, for example in RDF files, will be an important data-transmission format in the future. An important future project will be development of toolboxes that allow use of linked data in the data analysis and visualization systems used by USGS scientists. To start, the project user prototype application could be improved to directly link to additional data sources. For example, taxonomic data within the [Integrated Taxonomic Information System \(ITIS\)](#) could be referenced as a linked data source. Currently, taxonomic data are downloaded from ITIS and absorbed into the BioData and MARIS data sets. It would be preferable to link directly to ITIS resources to obtain the referential taxonomic data needed to describe species observations.

As government data providers, we have a responsibility to share our data with the widest audience possible, while ensuring that the data are usable, meaningful, and credible. The open government data movement is a natural extension of the [Transparency and Open Government Memorandum](#) and seeks to make government data freely available to the public in non-proprietary formats. Our investigation of semantic technologies can be used as a learning tool to help apply knowledge toward improving access to our USGS data assets through the use of linked data. For more detailed information on Linked Open Data, see Appendix 3. The CDI Semantic Web Working Group needs to further explore how semantic technologies such as linked open data can help us achieve these goals as we move into an era of greater government transparency and public collaboration.

Acknowledgments

The authors are grateful to Stephan Zednik of Rensselaer Polytechnic Institute, who patiently advised and taught the project team about use cases, the semantic development methodology, semantic technology, and much more. Many USGS employees and contractors contributed to the success of the project. In addition to the authors, the use case team included Alan Allwardt and Lisa Zolly. Brad Williams, Jeff Wendel, Mini Mathew, Julie Recker, Bruce Powell, and James Curry developed the prototype Web application, API, and system architecture. Lisa Zolly’s skill in English expression is greatly appreciated. David Mattli provided converted NHD data and ensured the functionality of the Center of Excellence for Geospatial Information Science SPARQL endpoint. Ariel Doumbova clarified many aspects of NHD data model of The National Map.

The project depended on data made available from the Multistate Aquatic Resources Information System (MARIS), which is a cooperative effort between state and federal agencies to share fisheries information collected as part of ongoing sampling programs. MARIS data are owned and provided by participating State natural-resource-management agencies, while technical support and hosting currently are provided by the [USGS Core Science Analytics and Synthesis program](#). Data were also made available from the Mineral Resources On-Line Spatial Data, maintained by the USGS Mineral

Resources program; the USGS Aquatic Bioassessment Data for the Nation (BioData); and the National Map of the United States, maintained by the USGS National Geospatial Program.

The project was supported financially by, and the project team owes its existence to, the USGS CDI. The CDI provides a forum for collaboration and brainstorming by bringing together expertise from external partners and representatives across the USGS who are involved in research, data management, and information technology. Through partnerships and working groups, the CDI leads the development of data-management tools and practices, cyber infrastructure, collaboration tools, and training in support of scientists and technology specialists throughout the project life cycle.

References Cited

- American National Standards Institute (ANSI), 1975, ANSI/X3/SPARC study group on Database Management Systems, interim report: FDT—Bulletin of ACM SIGMOD, v. 7, no. 2.
- Apache Jena, 2012a, Apache Jena: Apache Software Foundation, accessed November 8, 2012 at: http://jena.apache.org/about_jena/about.html
- Apache Jena, 2012b, Apache Jena— TDB: Apache Software Foundation, accessed November 8, 2012 at: <http://jena.apache.org/documentation/tdb/index.html>
- Apache Jena, 2012c, Fuseki—Deriving RDF data over HTTP: Apache Software Foundation, accessed November 8, 2012 at http://jena.apache.org/documentation/serving_data/index.html
- Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G., 2012, Turtle, terse RDF triple language: World Wide Web Consortium, accessed February, 12, 2013 at <http://www.w3.org/TR/turtle>.
- Berners-Lee, Tim, 2010, Open, linked data for a global community: World Wide Web Consortium, gov2.0 Expo, May 26, 2010, Washington Convention Center, Washington, D.C., Keynote Address, accessed November 3, 2014 at <http://youtu.be/ga1aSJXCFe0>.
- Cox, S., 2010, Geographic information—Observations and measurements, version 2.0.0: Open Geospatial Consortium, Inc., 46 p., accessed February, 6, 2013 at <http://www.opengeospatial.org/standards/om>. [OGC Abstract Specification Topic 20, Reference Number OGC 10-004r3 and ISO 19156.]
- Cox, S., 2011, OWL representation of ISO 19156 (Observation model): def.seegrid, accessed February, 6, 2013 at <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation>.
- Dublin Core, 2013, Dublin Core Metadata Initiative—Metadata innovation: Dublin Core Metadata Initiative, accessed February 4, 2013 at <http://dublincore.org>.
- Epimorphics, 2012, Elda 1.2.16 and implementation of the Linked Data API: Epimorphics Ltd., accessed November 8, 2012 at <http://elda.googlecode.com/hg/deliver-elda/src/main/webapp/lda-assets/docs/E1.2.16-advanced.html>
- Fox, P., and McGuinness, D.L., 2008, TWC Semantic Web Methodology: Tetherless World Constellation, accessed October 22, 2012 at http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology.
- Heath, Tom, 2012, Linked Data – Connect Distributed Data across the Web, accessed February 19, 2013 at <http://linkeddata.org>.
- Institute for Human and Machine Cognition (IHMC), 2012, IHMC Cmap Tools knowledge modeling kit: Pensacola, FL, Institute for Human and Machine Cognition, accessed August 7, 2012 at http://cmapdownload.ihmc.us/coe/Web_InstallersV5.0/install.htm.
- International Standards Organization (ISO), 2011, ISO/IEC 13249-3:2011—Information technology, Database languages, SQL multimedia and application packages, Part 3—Spatial (4th ed): International Standards Organization.

- Kifer, Michael, 2008, Rule interchange format—The framework, web reasoning and rule systems: Springer, Lecture Notes in Computer Science.
- Mattli, D., and Viers, W., 2012, Geospatial Semantics and Ontology: Center of Excellence for Geospatial Information Science, accessed August, 7, 2012 at <http://cegis.usgs.gov/ontology.html>.
- Moreau, Luc, Groth, Paul, Zhao, Jun, 2010, OPM Tutorial: Future Internet Symposium, Berlin, 20 September 2010, accessed August 15, 2012 at <http://openprovenance.org/tutorial/>.
- Multistate Aquatic Resources Information System (MARIS), 2012 Maris data: Multistate Aquatic Resources Information System, accessed August 6, 2012 at <http://marisdata.org>.
- National Fish Habitat Partnership (2d ed), 2012, National fish habitat action plan: Association of Fish and Wildlife Agencies, Washington D.C., 40 p., accessed October 10, 2012 at http://fishhabitat.org/sites/default/files/www/NFHP_AP_Final.pdf.
- Open Knowledge Foundation, 2014, The open definition: Open Knowledge, accessed November 3, 2014 at <http://opendefinition.org/>.
- Perry, M., and Herring, J., 2012, Open Geospatial Consortium GeoSPARQL—A Geographic Query Language for RDF Data, Open Geospatial Consortium Inc, v.1.0, accessed January 7, 2013 at <http://www.opengis.net/doc/IS/geosparql/1.0>.
- Schmachtenberg, Max, Bizer, Christian, Jentzch, Anja, and Cyganiak, Richard, 2014, The linking open data cloud diagram: Data Hub, accessed October 31, 2014 at <http://lod-cloud.net/>.
- Solid Earth and Environment GRID (SEE GRID), 2012, Observations and sampling—Solid earth and environment GRID: Solid Earth and Environment GRID, accessed October 10, 2012 at https://www.seegrid.csiro.au/wiki/AppSchemas/ObservationsAndSampling#Observation_Model.
- Simley, J., and Dombouya, A., 2012, National hydrography dataset—Linear referencing: U.S. Geological Survey Fact Sheet 2012–3068, 2 p., <http://pubs.usgs.gov/fs/2012/3068/>.
- U.S. Geological Survey, 2012a, BioData – Aquatic bioassessment data for the nation: U.S. Geological Survey, accessed August 7, 2012 at <https://aquatic.biodata.usgs.gov>.
- U.S. Geological Survey, 2012b, National Geochemical Survey, accessed December 7, 2012 at <http://tin.er.usgs.gov/geochem/>.
- U.S. Geological Survey, 2012c, National GAP Analysis Program (GAP)—Core Science Analytics and Synthesis: U.S. Geological Survey, accessed October 10, 2012 at <http://gapanalysis.usgs.gov>.
- U.S. Geological Survey, 2012d, Geographic information in the NGS Database: U.S. Geological Survey, accessed August, 7 2012 at <http://tin.er.usgs.gov/geochem/doc/geography.htm>.
- U.S. Geological Survey, 2012e, NHD tools: U.S. Geological Survey, accessed August 7, 2012 at <http://nhd.usgs.gov/tools.html#hem>.
- Wagner, A., Ladwig, G., Tran, T., 2011, Browsing-oriented semantic faceted search: Karlsruhe, Germany, Karlsruhe Institute of Technology, Institute of Applied Informatics and Formal Description Methods, accessed September 25, 2012 at <http://www.slideshare.net/ajwagner/browsingoriented-semantic-faceted-search>.
- World Wide Web Consortium, 2012a, Semantic web: World Wide Web Consortium,, 1 p., accessed October 11, 2012 at <http://www.w3.org/standards/semanticweb/>.
- World Wide Web Consortium, 2009, SKOS Simple knowledge organization system: World Wide Web Consortium, accessed October 11, 2012 at <http://www.w3.org/2004/02/skos/>.
- World Wide Web Consortium, 2012b, OWL web ontology language overview: World Wide Web Consortium, accessed October 11, 2012 at <http://www.w3.org/TR/owl-features/>.
- World Wide Web Consortium, 2004, SWRL—A semantic web rule language combining owl and RuleML: World Wide Web Consortium, accessed October 11, 2012 at <http://www.w3.org/Submission/SWRL/>.

World Wide Web Consortium, 2008, SPARQL query language for RDF: World Wide Web Consortium, accessed October 11, 2012 at <http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/>.

ISSN 2331-1258 (online)
<http://dx.doi.org/10.3133/ofr20151004>

Glossary

GeoSPARQL A geographic query language for RDF Data.

Linked Data A structured method of publishing data as resolvable Web resources identified using URIs and structured using RDF. Linked data can be easily read by computers and allows for cross-linking of data from different sources.

Linked Open Data Linked data that are openly licensed for reuse by anyone at no cost. See appendix 3 for more information.

Ontology A formal, explicit representation of vocabulary and taxonomy that allows automatic construction of relationships among objects and concepts.

Open Data Information that is available for anyone to use, for any purpose, at no cost

Provenance A record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for accessing authenticity, enabling trust, and allowing reproducibility. Provenance has also traditionally meant the historical record of data processing, its origin, and ownership.

Resource Description Framework (RDF) A standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”). More information at <http://www.w3.org/RDF/>.

RDFa A W3C Recommendation, Resource Description Framework in Attributes (RDFa) adds a set of attribute-level extensions to HTML, XHTML, and various XML-based document types for embedding rich metadata within Web documents. More information at <http://www.w3.org/TR/xhtml-rdfa-primer/>.

RDFS An extension to RDF that provides classes and properties to improve descriptions of vocabularies. RDFS provides a rudimentary representation of ontologies.

Representational State Transfer (REST) An architectural style that uses a stateless, client-server, cacheable communications protocol, such as HTTP, to make calls between machines. RESTful refers to an application that uses REST. (<http://rest.elkstein.org/>)

SPARQL Protocol and RDF Query Language (SPARQL) is a RDF query language. See appendix 1 for more information.

SPARQL Endpoint a conformant SPARQL protocol service used to query a knowledge base via the SPARQL language with results typically returned in one or more machine-processable formats.

Semantic Web A framework of formats, technologies, and resources enabling structured Web content that includes explicit definition of the meaning and relationships among information elements. The goal of Semantic Web is to transmit data that can be processed directly by computers. More information at <http://www.w3.org/standards/semanticweb/>.

Triple An RDF statement consisting of subject, predicate, and object.

Triple store A database for the storage and retrieval of RDF triples and graphs.

Uniform Resource Identifier (URI) A string of characters used for identifying, or naming, a resource. When the URI is in the form of a network address (uniform resource locator, URL) it can allow direct navigation to either the item or to information about the item.

Uniform Resource Locator (URL) A type of URI that identifies the global address of a document or resource on the Web.

Use Case A document that describes the interactions between external actors and the system under consideration to accomplish a goal.

Web Ontology Language OWL A family of specifications for expressing ontologies using RDF. More information at <http://www.w3.org/TR/owl-ref/>.

Appendix 1: Semantic Web Technologies (Overview)

According to the World Wide Web Consortium (W3C):

In addition to the classic “Web of documents” W3C is helping to build a technology stack to support a “Web of data,” the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term “Semantic Web” refers to W3C’s vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS. (World Wide Web Consortium, 2012a)

Data Interchange

The main goal of the Semantic Web is to create a Web of data enabling computers to do useful work through systems that can interact and exchange data in a meaningful way. The following sections describe parts of the semantic web technical stack that enable this interchange of data.

Resource Description Framework

A standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”). More information at <http://www.w3.org/RDF/>. See Appendix 2 for more detailed information on RDF.

Resource Description Framework in Attributes (RDFa)

A W3C Recommendation, Resource Description Framework in Attributes (RDFa) adds a set of attribute-level extensions to HTML, XHTML, and various XML-based document types for embedding rich metadata within Web documents. More information at <http://www.w3.org/TR/xhtml-rdfa-primer/>.

Linked Data

A structured method of publishing data as resolvable Web resources identified using URIs and structured using RDF. Linked data can be easily read by computers and allows for cross-linking of data from different sources (Heath, 2012). Appendix 3 contains a more detailed description of Linked Data and further information can be found at <http://linkeddata.org>.

Vocabularies

Vocabulary languages are an integral component of the Semantic Web, providing a formal mechanism for knowledge representation and inference.

RDF Schema (RDFS)

The RDF Vocabulary Description Language (RDF Schema) is a language that provides a mechanism to describe how to use RDF to describe RDF vocabularies. RDFS provides classes and properties that may be used to describe classes, properties, and other resources. RDFS is a W3C Recommendation. For more information on RDFS, see <http://www.w3.org/TR/rdf-schema/>.

Simple Knowledge Organization Systems (SKOS)

Simple Knowledge Organization System (SKOS) is a family of formal languages designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured, controlled vocabulary (World Wide Web Consortium, 2009). SKOS is built upon RDF and RDFS. Its main objective is to enable easy publication of controlled structured vocabularies for the Semantic Web. For more information on SKOS see <http://www.w3.org/TR/skos-primer/>.

Web Ontology Language (OWL)

The Web Ontology Language (OWL) builds upon the capabilities of RDF and RDFS to increase machine interoperability by explicitly defining the meaning of terminology used to describe data on the Web. The Web Ontology Language provides formal semantics that enable machines to perform reasoning tasks on data for automated processing and integration. (World Wide Web Consortium, 2012b) OWL and OWL2 are used to refer to the 2004 and 2009 specifications, respectively.

Query

The SPARQL Protocol and RDF Query Language (SPARQL) is an RDF query language and the official query language of the Semantic Web (World Wide Web Consortium, 2008). SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be results sets or RDF graphs. A SPARQL query can also be used to construct and return a new RDF graph from the source RDF based on a transformation defined in the query. SPARQL has a SQL-like syntax, as seen in the following example.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
  ?person a foaf:Person.
  ?person foaf:name ?name.
  OPTIONAL { ?person foaf:mbox ?email }
}
```

SPARQL 1.1 introduced update (write) and federated query support. SPARQL is a W3C Recommendation. For more detailed information on SPARQL see <http://www.w3.org/TR/rdf-sparql-query/>.

Rules

Rule systems provide a formal mechanism to describe how new information can be inferred based on a possibly very complex chain of antecedent conditions. A very simple example of a rule would be “my father’s brother is my uncle.” Rules can become very complex and contain conditions that test and compare values or infer the existence of otherwise unknown resources. Most rule systems and inference axioms in the Semantic Web are based on Description Logic. Two important technologies for Rules in the Semantic Web are the Rule Interchange Format (RIF) and the Semantic Web Rule Language (SWRL).

Rule Interchange Format

The Rule Interchange Format (RIF) is a W3C Recommendation for an interchange format for rules in the Semantic Web. Although originally envisioned by many as a “rules layer” for the Semantic Web, in reality, the design of RIF is based on the observation that there are many “rules languages” in existence, and what is needed is to exchange rules between them (Kifer, 2008). For more information on RIF see, <http://www.w3.org/TR/rif-primer/>.

Semantic Web Rule Language

The Semantic Web Rule Language (SWRL) is a W3C Submission based on a combination of the OWL DL and OWL Lite sublanguages of the OWL Web Ontology Language with the Unary/Binary Datalog RuleML sublanguages of the Rule Markup Language. (World Wide Web Consortium, 2004) Rules are of the form of an implication between an antecedent (body) and consequent (head). The intended meaning can be read as: whenever the conditions specified in the antecedent hold, then the conditions specified in the consequent must also hold. (World Wide Web Consortium, 2004) SWRL is currently not a W3C Recommendation. For more information see <http://www.w3.org/Submission/SWRL/>.

Appendix 2: Introduction to Resource Description Framework

Resource Description Framework

Resource Description Framework (RDF) is the cornerstone of the Semantic Web technology stack and is used for data interchange on the Web. This appendix describes the basic concepts of the RDF model and the most popular RDF serialization formats.

RDF Data Model

(RDF) is the W3C standard model for data interchange on the Web (www.w3.org/RDF). RDF provides a structure for describing resources and relationships between resources, all of which are represented by Uniform Resource Identifiers (URIs). The way the RDF model organizes data is in the form of statements, often referred to as “triples” because each statement consists of three parts: a subject, a predicate, and an object. This structure follows the grammar rules of an English sentence, with the subject being the “thing” that carries out an action, the predicate tells us about the “action,” and the object is the other “thing” upon which the action is done.

For example, take the statement “The fish swims in the lake.” The subject would be the fish, the predicate would be “swims in,” and the object would be the lake. In RDF, this would be represented in a graph showing the relationship between the fish and the lake as follows:

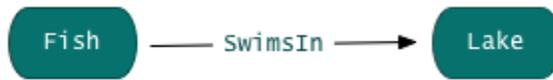


Figure 2-1. Graph depicting the triple where the subject is Fish, the predicate is Swims In and the Object is Lake.

RDF uses URIs to identify the subject, predicate, and the object unless the object is a literal value. In our example, the subject or “Fish” could be identified as a resource with a URI such as: “http://example.org/fishes#fish.” The predicate “Swims In” could be identified with another URI such as: “http://example.org/fishes#swimsIn.” The object “Lake” could either be represented as a URI such as: “http://example.org/lakes#lake,” or it could be represented by a literal value such as “Lake”^{^^xsd:string}.

Part of the power behind the RDF model is the ability to connect statements together. To continue our example, let’s add another statement: “The lake is named Clear Lake.” This triple could be visualized as follows:



Figure 2-2. Graph depicting the triple where the Subject is Lake, the predicate is Named and the object is Clear Lake.

We can now connect these two unique statements, or triples, together, creating a larger graph.

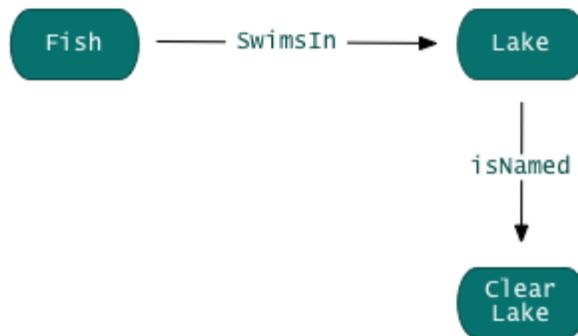


Figure 2-3. A graph depicting two triples. The first triple has the Subject Fish, the predicate SwimsIn with the Object Lake. The second triple has the Subject Lake, the predicate is Named, and the Object Clear Lake.

As you can see from the example, RDF is represented through a graph of subjects, objects, and their relationships to each other each represented by a URI. This structure allows the RDF data structure great flexibility in connecting to other data on the Web without being dependent on a highly structured schema.

RDF Serializations

RDF is a data model that can be serialized in multiple formats. Serialization allows the encoding of a data model into a format that computers can read, store, and process. The most common

serialization formats used for RDF data are RDF/XML, Terse RDF Triple Language (Turtle), and N-triples. They can all represent the same RDF data model, but each does this with its own syntax. The same RDF graph can be serialized in the RDF/XML, Turtle, and N-triple formats and interpreted by a computer as semantically and structurally identical.



Figure 2-4. Graphical representation of the triple where the subject is Fish, the predicate is SwimsIn, and the Object is Lake.

RDF/XML

RDF/XML is an XML syntax for expressing an RDF graph and is the W3C standard for RDF serialization.

```
<rdf:Description rdf:about="#fish">
  <rdf:type rdf:resource="#Animal"/>
  <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">fish</rdfs:label>
  <swimsIn rdf:resource="#lake"/>
</rdf:Description>
```

N-Triples

N-Triples is a plain text format that allows one triple per line and describes resources with the unabbreviated URI. It was primarily designed for ease of machine parsing and generation, but not necessarily for ease of human readability.

```
<http://example.org/example#fish> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>↔
<http://example.org/example#Animal> .
<http://example.org/example#fish> <http://www.w3.org/2000/01/rdf-schema#label>↔
"fish"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://example.org/example#fish> <http://example.org/example#swimsIn>
<http://example.org/example#lake> .
```

Terse RDF Triple Language (Turtle)

Terse RDF Triple Language (Turtle) is a simple, compact syntax for expressing RDF triples in a way that resembles natural language.

```
:fish
  rdf:type      :Animal ;
  rdfs:label "fish"^^xsd:string ;
  :swimsIn :lake .
```

RDF Resources

W3School's RDF Tutorial

<http://www.w3schools.com/rdf/default.asp>

W3C's description of the RDF Semantic Web Standard
<http://www.w3.org/RDF/>

Cambridge Semantics' Introduction to RDF
<http://www.cambridgesemantics.com/semantic-university/rdf-101>

Appendix 3: Linked Open Data

This appendix will explain the concepts of open data, linked data, and linked open data.

Open Data

There is an international movement towards increasing global knowledge through the opening of data. The Open Knowledge Foundation defines Open Data as data that can be “freely used, modified, and shared by anyone for any purpose” (Open Knowledge Foundation, 2014).

Initiatives, such as the [Transparency and Open Government Memorandum](#) and the [Open Government Partnership](#) have led to expectations that government data be made open for public use. Government Data holdings are now more widely accessible through catalog systems such as [data.gov](#) and [data.gov.uk](#). Semantic Web technologies such as Linked Data can catalyze data openness by providing standardized mechanisms for data access, distribution, integration and reuse on the Web.

Linked Data

Linked Data is the practice of linking together data on the Web through RDF and URIs, making discovery of related data easier. Linked data isn't necessarily open data, however. Linked data techniques can be used in closed systems using proprietary data.

Linked Open Data

You can think of Linked Open Data as the union of Open Data and Linked Data. Linked Open Data ties together the Web of data by creating relationships between different open data sources, thereby easing data discovery. A good visualization of interlinked open data sources, is the “Linking Open Data cloud diagram” shown in figure 3-1.

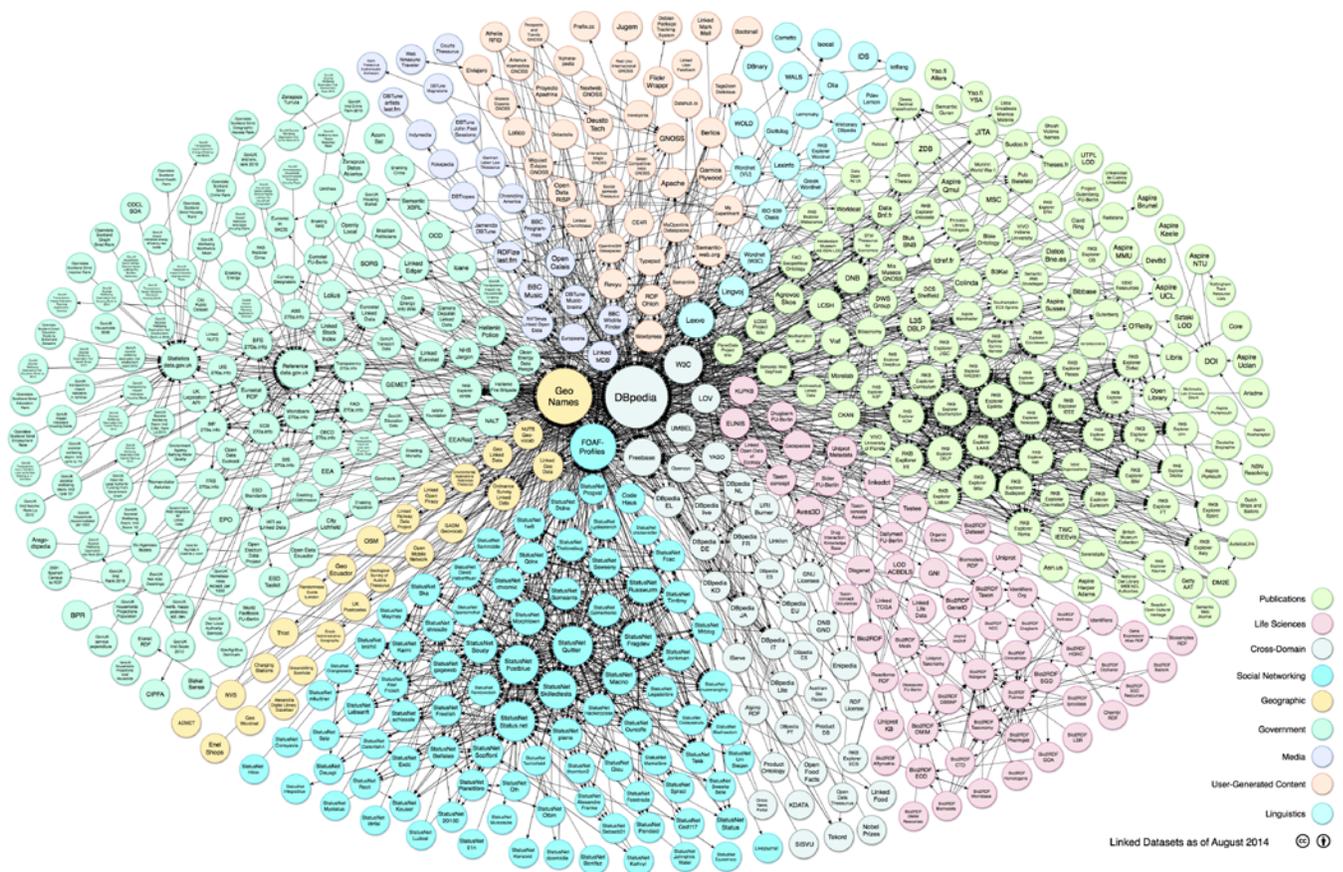


Figure 3-1. Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net/> (Schmachtenberg, 2014) This visualization shows the linkages between different LOD sources and has been color coded by different domains including: Publications; Life Sciences; Cross-Domain; Social Networking; Geographic; Government; Media; User-Generated Content; and Linguistics. For closer inspection of the diagram see http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.png.

How can we as government data stewards create LOD sources for others to use? Tim Berners-Lee introduced a roadmap for creating Open Government Data (OGD) with his idea of **5 star Data** (Berners-Lee, 2010). His rating system gives data stewards incremental goals to strive towards, with the ultimate five-star rating being given for those who achieve the state of Linked Open Data that include references to other data sources. This series of steps is described in Appendix table 3-1.

Table 3-1. Description of Tim Berners-Lee’s five star data model. (CSV, comma-separated values; URI, universal resource indicator)

Rating Earned	Explanation	Badge
Earn 1 Star if:	your data are available on the Web (whatever format) with an open license	
Earn 2 Stars if:	your data are available as machine-readable structured data (for example, Excel instead of image scan of a table)	
Earn 3 Stars if:	your data are in non-proprietary formats (for example, CSV instead of Excel)	
Earn 4 Stars if:	you use URIs to identify things, so that people can point at your stuff	
Earn 5 Stars if:	you link your data to other data to provide context	

In order to achieve a four-star rating, your data must be accessible through a URI. To get a five-star rating, your data must link to other data sources. These ratings can be achieved through Linked Open Data technologies.

Achieving a five-star data rating can seem overwhelming if you are just starting out, but it can be achieved through a series of simple steps. Figure 3-2 below shows how data owners can make progress through a series of small steps to make their data more usable.

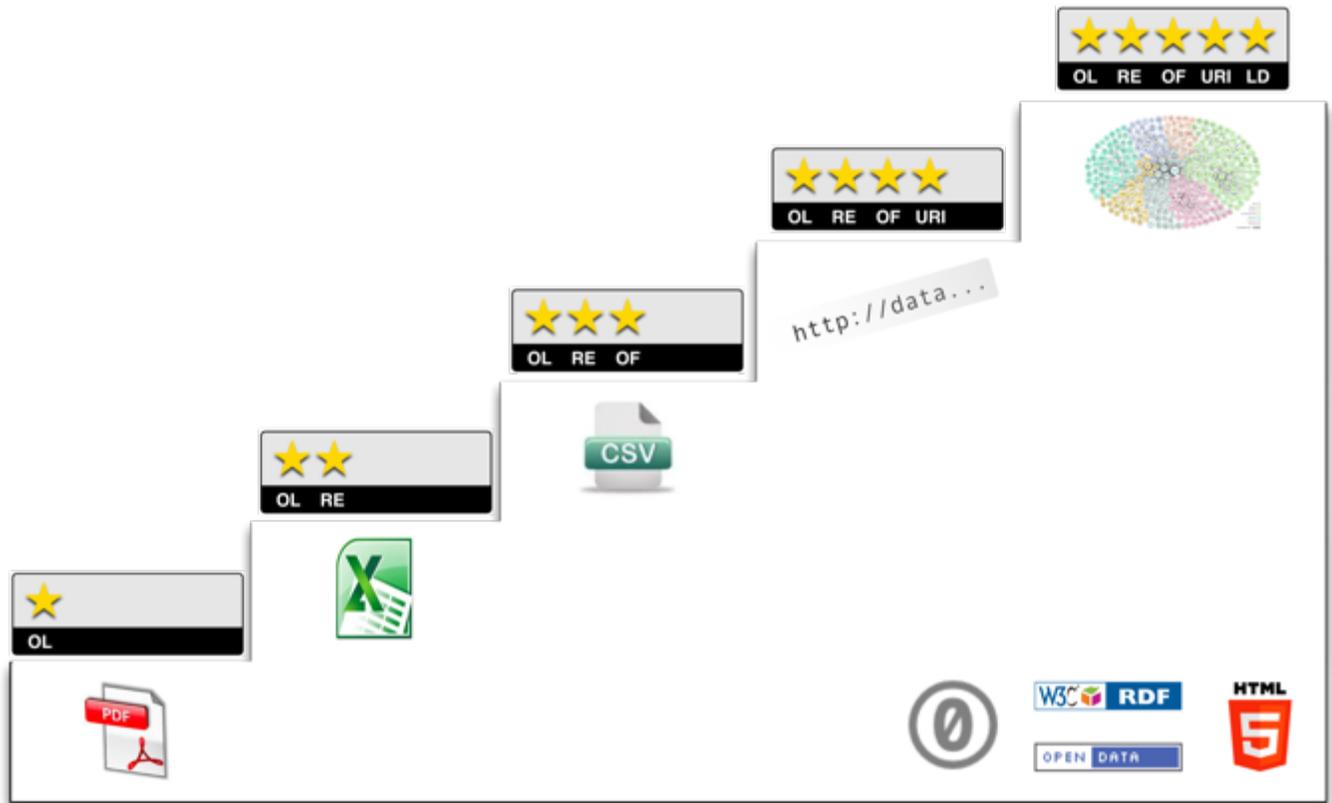


Figure 3-2. A visualization of the steps needed to achieve a five-star data rating, provided by <http://5stardata.info/>.

Linked Open Data Resources

World Wide Web Consortium’s (W3C) Overview on Linked Data
<http://www.w3.org/standards/semanticweb/data>

Tim Berners-Lee’s Linked Data Design Rules
<http://www.w3.org/DesignIssues/LinkedData.html>

Linked Data: Evolving the Web into a Global Data Space—An online book giving an overview of Linked Data
<http://linkeddatabook.com/editions/1.0/>

Linked Open Data: The Essentials—An online book about the importance of Linked Open Data for government data
www.semantic-web.at/LOD-TheEssentials.pdf

For all things Linked Data
<http://linkeddata.org/>

Appendix 4: API SPARQL Statements

SELECT Statement 1: is sent to the MARIS endpoint to query for all water quality observations found within the Beechwood Lake sampling site.

Parameters Passed: huc=02050104&wq_sampling_site=Beechwood%20Lake

```
PREFIX usgs: <http://www1.usgs.gov/linkedata/usgs#>
PREFIX sam: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>
PREFIX om: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX nure-site: <http://mrddata.usgs.gov/nuresed/site/>
PREFIX conc: <http://mrddata.usgs.gov/geochem/concentration#>
PREFIX method: <http://mrddata.usgs.gov/geochem/method#>
PREFIX geo: <http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX basic: <http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#>
PREFIX usgs: <http://www1.usgs.gov/linkedata/usgs#>
PREFIX sam: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>
PREFIX om: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX nure-site: <http://mrddata.usgs.gov/nuresed/site/>
PREFIX conc: <http://mrddata.usgs.gov/geochem/concentration#>
PREFIX method: <http://mrddata.usgs.gov/geochem/method#>
PREFIX geo: <http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX basic: <http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#>
```

```
SELECT (GROUP_CONCAT(DISTINCT ?obsId; SEPARATOR=",") AS ?wq_list)
{
  ?observation sam:samplingLocation ?site ;
    sam:relatedObservation ?waterObs ;
    sam:samplingTime ?samplingDateId .
  ?waterObs om:observedProperty ?waterProp .
  ?waterProp rdfs:label ?characteristic .
  ?samplingDateId usgs:begin ?samplingDate .
  ?site geo:within ?hucId ;
    usgs:waterName ?samplingSite .
  FILTER (regex(str(?hucId), "hucf/", "i" ))
  FILTER (regex(str(?observation), "water", "i" ))
  FILTER (regex(str(?huc), "^02050104", "i"))
  FILTER (?samplingSite IN ("Beechwood Lake"))
  BIND (strafter(str(?hucId), "hucf/") AS ?huc )
  BIND (strafter(str(?observation), "samplingcollection/") AS ?obsId)
```

}

SELECT Statement 1 Returns: the list of water quality resources from the MARIS endpoint that met the filter criteria.

<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1191>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1194>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1193>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1192>

SELECT Statement 2: is sent to the MARIS endpoint to query for all species observations found at the Beechwood Lake sampling site that include the species *Esox niger* (ITIS TSN 162143)

Parameters Passed: huc=02050104&so_tsn=162143&so_sampling_site=Beechwood%20Lake

PREFIX usgs: <<http://www1.usgs.gov/linkedata/usgs#>>
PREFIX sam: <<http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>>
PREFIX om: <<http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>>
PREFIX dc: <<http://purl.org/dc/elements/1.1/>>
PREFIX nure-site: <<http://mrdata.usgs.gov/nuresed/site/>>
PREFIX conc: <<http://mrdata.usgs.gov/geochem/concentration#>>
PREFIX method: <<http://mrdata.usgs.gov/geochem/method#>>
PREFIX geo: <<http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/>>
PREFIX rdfs: <<http://www.w3.org/2000/01/rdf-schema#>>
PREFIX xsd: <<http://www.w3.org/2001/XMLSchema#>>
PREFIX basic: <<http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#>>

SELECT (GROUP_CONCAT(DISTINCT ?obsId; SEPARATOR=",") AS ?wq_list)

```
{
?observation sam:samplingLocation ?site ;
              sam:samplingMethod ?method ;
              sam:samplingTime ?samplingDateId ;
              usgs:itisTsn ?tsn .
?tsn usgs:scientificName ?tsnName .
?method rdfs:label ?methodLabel .
?samplingDateId usgs:begin ?samplingDate .
?site geo:within ?hucId ;
       usgs:waterName ?siteId .
FILTER (regex(str(?hucId), "hucf/", "i" ))
FILTER (regex(str(?huc), "^02050104", "i"))
FILTER (?tsnId IN ("162143"))
FILTER (?siteId IN ("Beechwood Lake"))
```

```
    BIND (strafter(str(?hucId),"hucf/") AS ?huc )
    BIND (strafter(str(?tsn),"tsn/") AS ?tsnId )
    BIND (strafter(str(?observation),"samplingcollection/") AS ?obsId)
  }
```

SELECT Statement 2 Returns: the list of species observation resources from the MARIS endpoint that met the filter criteria

```
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA203098>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA45318>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA69275>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA76533>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA194881>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA105912>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA51818>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA61166>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA70602>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA639>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA164744>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA96144>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA71568>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA123332>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA195843>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA161191>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA211182>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA182193>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA113795>,
```

<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA26528>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA55101>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA200473>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA82621>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA139559>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA19812>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA151059>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA182954>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA11891>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA76971>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA161944>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA153613>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA216828>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA80690>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA179263>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA153723>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA91260>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA192014>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA165478>,
http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA151664

SELECT Statement 3: is sent to the NGS endpoint to query for all lead (Pb) observations found within HUC 02050104

Parameters Passed: huc=02050104&sg_chemical_species=conc:Pb

PREFIX usgs: <<http://www1.usgs.gov/linkedata/usgs#>>

PREFIX sam: <<http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>>
 PREFIX om: <<http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>>
 PREFIX dc: <<http://purl.org/dc/elements/1.1/>>
 PREFIX nure-site: <<http://mrdata.usgs.gov/nuresed/site/>>
 PREFIX conc: <<http://mrdata.usgs.gov/geochem/concentration#>>
 PREFIX method: <<http://mrdata.usgs.gov/geochem/method#>>
 PREFIX geo: <<http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/>>
 PREFIX rdfs: <<http://www.w3.org/2000/01/rdf-schema#>>
 PREFIX xsd: <<http://www.w3.org/2001/XMLSchema#>>
 PREFIX basic: <<http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#>>

```
SELECT (GROUP_CONCAT(DISTINCT ?obsId; SEPARATOR=",") AS ?sg_list)
{
  ?observation om:featureOfInterest ?sample ;
    om:observedProperty ?element ;
    om:procedure ?methodBase .
  ?sample sam:samplingLocation ?site ;
    sam:samplingTime ?collectiondate .
  ?site usgs:huc ?huc ;
    dc:identifier ?siteBase .
  ?collectiondate usgs:date ?date .
  FILTER (regex(str(?huc), "^2050104", "i"))
  FILTER (?element IN (conc:Pb))
  BIND (strafter(str(?observation),"observation/") AS ?obsId)
}
```

SELECT Statement 3 Returns: The list of resources from the NGS endpoint that met the filter criteria.

<http://mrdata.usgs.gov/geochem/observation/C-262541/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-157189/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-156799/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-163047/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-157036/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-162846/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-280309/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-157012/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-156951/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-156527/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-156873/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-157080/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-163605/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-156626/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-157183/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-210931/pb_icp40>,
 <http://mrdata.usgs.gov/geochem/observation/C-262598/pb_icp40>

SELECT Statement 4: is sent to the BioData endpoint to query for all species observations found at the Beechwood Lake sampling site that include the species *Esox niger* (ITIS TSN 162143)
Parameters: huc=02050104&so_tsn=162143&so_sampling_site=Beechwood%20Lake

PREFIX usgs: <http://www1.usgs.gov/linkedata/usgs#>
PREFIX sam: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>
PREFIX om: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX nure-site: <http://mrddata.usgs.gov/nuresed/site/>
PREFIX conc: <http://mrddata.usgs.gov/geochem/concentration#>
PREFIX method: <http://mrddata.usgs.gov/geochem/method#>
PREFIX geo: <http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX basic: <http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#>

```
SELECT (GROUP_CONCAT(DISTINCT ?obsId; SEPARATOR=",") AS ?so_list)
{
  ?observation om:featureOfInterest ?sample ;
    om:observedProperty ?abundance .
  ?abundance usgs:taxon ?taxon .
  ?taxon usgs:tsn ?tsnId ;
    usgs:biodataTaxon ?bioTaxon .
  ?bioTaxon usgs:name ?tsnName .
  ?sample sam:samplingLocation ?site;
    sam:samplingTime ?time ;
    usgs:samplingMethod ?samplingMethod .
  ?site geo:within ?hucId ;
    usgs:siteNumber ?siteId ;
    rdfs:label ?siteName .
  ?time usgs:year ?year ;
    usgs:date ?samplingDate .
  ?samplingMethod usgs:subMethod ?subMethod .
  ?subMethod usgs:gearUsed ?gear .
  ?gear rdfs:label ?methodLabel .
  FILTER (regex(str(?hucId), "huc/", "i" ))
  FILTER (regex(str(?huc), "^02050104", "i"))
  FILTER (?tsnId IN ("162143"))
  FILTER (?siteId IN ("Beechwood Lake"))
  BIND (strafter(str(?hucId),"huc/") AS ?huc )
  BIND (strafter(str(?observation),"obs/") AS ?obsId)
}
```

Query 4 Returns: No observations were found so this query returns an empty list.

All of the select queries have returned the observation resource lists, which will now be used in SPARQL CONSTRUCT statements that will generate new RDF.
CONSTRUCT Statement 1: uses the returned lists from SELECT Statements 1 & 2 as a filter to Query the MARIS endpoint and return results as RDF.

```
PREFIX usgs: <http://www1.usgs.gov/linkedata/usgs#>
PREFIX sam: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>
PREFIX om: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX nure-site: <http://mrdata.usgs.gov/nuresed/site/>
PREFIX conc: <http://mrdata.usgs.gov/geochem/concentration#>
PREFIX method: <http://mrdata.usgs.gov/geochem/method#>
PREFIX geo: <http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX basic: <http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#>
```

CONSTRUCT

```
{
?observation sam:samplingLocation ?site ;
    sam:relatedObservation ?waterObs ;
    sam:samplingTime ?samplingDate ;
    sam:samplingMethod ?method ;
    usgs:marisId ?marisId ;
    usgs:waterId ?waterId ;
    usgs:itisTsn ?tsn ;
    usgs:effortTime ?effortTime ;
    usgs:stateSpeciesId ?stateSpeciesId ;
    usgs:targetStandard ?targetStandard .
?waterObs om:observedProperty ?waterProp ;
    om:result ?result .
?waterProp rdfs:label ?characteristic .
?result basic:number ?number ;
    basic:unit ?unit .
?samplingDate usgs:begin ?samplingDateBegin ;
    usgs:end ?samplingDateEnd . ?site geo:within ?hucId ;
    usgs:waterName ?samplingSite ;
    usgs:waterType ?waterType ;
    usgs:collectionLocationType ?locationType ;
    usgs:collectionAccuracyDescription ?accDesc ;
    usgs:latitude ?lat ;
    usgs:longitude ?long .
?lat basic:number ?number1 ;
    basic:unit ?unit1 .
```

```

?long basic:number ?number3 ;
    basic:unit ?unit3 .
?tsn usgs:scientificName ?tsnName ;
    usgs:commonName ?cn .
?method rdfs:label ?methodLabel .
?effortTime basic:number ?number4 ;
    basic:unit ?unit4 .
}
WHERE
{
?observation sam:samplingLocation ?site ;
    sam:relatedObservation ?waterObs ;
    sam:samplingTime ?samplingDate ;
    usgs:marisId ?marisId ;
    usgs:waterId ?waterId .
?waterObs om:observedProperty ?waterProp ;
    om:result ?result .
?waterProp rdfs:label ?characteristic .
?result basic:number ?number ;
    basic:unit ?unit .
?samplingDate usgs:begin ?samplingDateBegin ;
    usgs:end ?samplingDateEnd .
?site geo:within ?hucId ;
    usgs:waterName ?samplingSite ;
    usgs:waterType ?waterType ;
    usgs:latitude ?lat ;
    usgs:longitude ?long .
?lat basic:number ?number1 ;
    basic:unit ?unit1 .
?long basic:number ?number3 ;
    basic:unit ?unit3 .
OPTIONAL
{
    ?observation sam:samplingMethod ?method .
    ?method rdfs:label ?methodLabel .
}
OPTIONAL
{
    ?observation usgs:itisTsn ?tsn .
    ?tsn usgs:scientificName ?tsnName ;
    usgs:commonName ?cn .
}
OPTIONAL
{
    ?observation usgs:effortTime ?effortTime .
    ?effortTime basic:number ?number4 ;

```

```

    basic:unit ?unit4 .
}
OPTIONAL { ?observation usgs:stateSpeciesId ?stateSpeciesId . }
OPTIONAL { ?observation usgs:targetStandard ?targetStandard . }
OPTIONAL { ?site usgs:collectionLocationType ?locationType . }
OPTIONAL { ?site usgs:collectionAccuracyDescription ?accDesc . }
FILTER ( ?observation IN
(<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
203098>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
45318>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
69275>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
76533>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
194881>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
105912>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
51818>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
61166>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
70602>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
639>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
164744>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
96144>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
71568>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
123332>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
195843>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
161191>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
211182>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
182193>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA
113795>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA

```

26528>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA55101>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA200473>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA82621>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA139559>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA19812>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA151059>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA182954>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA11891>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA76971>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA161944>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA153613>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA216828>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA80690>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA179263>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA153723>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA91260>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA192014>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA165478>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/fish_PA151664>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1191>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1194>,
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_PA1193>,

```
<http://www1.usgs.gov/linkedata/swwg/maris/samplingcollection/PAI415117773045/sample/water_P
A1192>) )
}
```

CONSTRUCT Statement 2: uses the list returned in SELECT statement 3 as a filter used to query the NGS endpoint and return results as RDF.

```
PREFIX usgs: <http://www1.usgs.gov/linkedata/usgs#>
PREFIX sam: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>
PREFIX om: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX nure-site: <http://mrdata.usgs.gov/nuresed/site/>
PREFIX conc: <http://mrdata.usgs.gov/geochem/concentration#>
PREFIX method: <http://mrdata.usgs.gov/geochem/method#>
PREFIX geo: <http://www.opengis.net/ont/OGC-GeoSPARQL/1.0/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX basic: <http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#>
PREFIX ngs: <http://mrdata.usgs.gov/geochem/field#>
```

CONSTRUCT

```
{
?observation om:featureOfInterest ?sample ;
    om:observedProperty ?element ;
    om:procedure ?methodBase ;
    dcterms:isPartOf ?isPartOf ;
    om:result ?result .
?result basic:number ?basicNumber ;
    basic:unit ?basicUnit .
?sample sam:materialClass ?materialClass ;
    sam:samplingLocation ?site ;
    sam:samplingTime ?collectiondate ;
    ngs:collector ?collector ;
    ngs:descript ?desc ;
    ngs:dried ?dried ;
    ngs:medium ?medium ;
    ngs:source ?source ;
    ngs:source_mod ?sorcemod ;
    ngs:stype ?stype ;
    ngs:upsieve ?upsieve ;
    dc:identifier ?dcid ;
    ngs:grabs ?grabs ;
    ngs:grainsize ?grainsize ;
    ngs:smpgrsize ?smpgrsize .
?site ngs:acchanwid ?acchanwid ;
```

```

    ngs:flowrate ?flowrate ;
    ngs:flowstage ?flowstage ;
    ngs:strbed ?strbed ;
    ngs:veg ?veg ;
    ngs:watcol ?watcol ;
    ngs:waterdep ?waterdep ;
    dc:identifier ?siteBase ;
    geo:within ?within ;
    usgs:datum ?datum ;
    usgs:huc ?huc ;
    usgs:latitude ?lat ;
    usgs:longitude ?long ;
    ngs:contampot ?contampot ;
    ngs:contamsou ?contamsou ;
    ngs:fldplnwid ?fldplnwid ;
    ngs:photos ?photos ;
    ngs:setting ?setting .
?collectiondate usgs:date ?date .
}
WHERE
{
?observation om:featureOfInterest ?sample ;
    om:observedProperty ?element ;
    om:procedure ?methodBase ;
    dcterms:isPartOf ?isPartOf .
OPTIONAL
{
    ?observation om:result ?result .
    ?result basic:number ?basicNumber ;
    basic:unit ?basicUnit .
}
?sample sam:materialClass ?materialClass ;
    sam:samplingLocation ?site ;
    sam:samplingTime ?collectiondate ;
    ngs:collector ?collector ;
    ngs:descript ?desc ;
    ngs:dried ?dried ;
    ngs:medium ?medium ;
    ngs:source ?source ;
    ngs:source_mod ?sorcemod ;
    ngs:stype ?stype ;
    ngs:upsieve ?upsieve ;
    dc:identifier ?dcid .
OPTIONAL { ?sample ngs:grabs ?grabs . }
OPTIONAL { ?sample ngs:grainsize ?grainsize . }
OPTIONAL { ?sample ngs:smpgrsize ?smpgrsize . }

```

```

?site ngs:acchanwid ?acchanwid ;
    ngs:flowrate ?flowrate ;
    ngs:flowstage ?flowstage ;
    ngs:strbed ?strbed ;
    ngs:veg ?veg ;
    ngs:watcol ?watcol ;
    ngs:waterdep ?waterdep ;
    dc:identifier ?siteBase ;
    geo:within ?within ;
    usgs:datum ?datum ;
    usgs:huc ?huc ;
    usgs:latitude ?lat ;
    usgs:longitude ?long .
OPTIONAL { ?site ngs:contampot ?contampot . }
OPTIONAL { ?site ngs:contamsou ?contamsou . }
OPTIONAL { ?site ngs:fldplnwid ?fldplnwid . }
OPTIONAL { ?site ngs:photos ?photos . }
OPTIONAL { ?site ngs:setting ?setting . }
?collectiondate usgs:date ?date .
FILTER ( ?observation IN (<http://mrdata.usgs.gov/geochem/observation/C-262541/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-157189/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-156799/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-163047/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-157036/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-162846/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-280309/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-157012/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-156951/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-156527/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-156873/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-157080/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-163605/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-156626/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-157183/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-210931/pb_icp40>,
<http://mrdata.usgs.gov/geochem/observation/C-262598/pb_icp40> ) )
}

```