Earth Sci Inform (2015) 8:649–661
DOI 10.1007/s12145-014-0175-1

METHODOLOGY ARTICLE

A semantically enabled metadata repository for scientific data

Anne Wilson · Michael Cox · Don Elsborg · Doug Lindholm · Tyler Traver

Received: 30 March 2014 / Accepted: 19 August 2014 / Published online: 31 August 2014

Abstract The LASP Extended Metadata Repository (LEMR) is a semantically enabled repository of information (metadata) about the scientific datasets that LASP offers to the public. The repository enables the provision of consistent, current, verified metadata to our users. It serves as a Single Source of Truth for this information, enabling more rigorous metadata management and addressing problems related to duplication of information. The linked open data aspect of the repository allows interlinking of concepts both within and across organizations and web sites. Associated interfaces allow users to browse and search the metadata. This information can be dynamically incorporated into web pages, so web page content is always up-to-date and consistent across the lab. With this information we can generate metadata records in a variety of schemas, such as ISO or SPASE, allowing federation with other organizations interested in our data. We leveraged open source technologies to build the repository and the dynamic web pages that read from it. VIVO, an open source semantic web application, provided key capabilities such as ontology and triple store management interfaces. AngularJS, an open source JavaScript framework for building web dynamic applications, was also invaluable in developing web pages that provide semantically enabled public interfaces to the metadata. In this paper we discuss our use of these tools and what we had to craft in order to meet our lab-specific needs.

Keywords Semantic · Dynamic · Metadata · Repository · Ontology · Infrastructure · eScience

Communicated by: H. A. Babaie

A. Wilson (✉) · M. Cox · D. Elsborg · D. Lindholm · T. Traver
Laboratory for Atmospheric and Space Physics, University of Colorado at Boulder, 1234 Innovation Drive, Boulder, CO 80304, USA
e-mail: anne.wilson@lasp.colorado.edu

Introduction, motivation

Semantic web technologies are gaining traction in a variety of areas, including the earth sciences. However, these technologies are still in their early stages of development and use. While some data managers and providers offer semantic capabilities (for example, McGuinness et al. 2007, Narock and King 2008; Merka et al. 2008a; Narock et al. 2009; Fox et al. 2009; Narock and Fox 2012, Arko 2014), their number is relatively small. That may be due to a perceived high 'barrier to entry' in enabling semantic capabilities. Tools, instructions, and help have not been readily available for the general practitioner in data management. We report here on our experience building a semantically enabled metadata repository and associated services, leveraging open source tools to do so.

The Laboratory for Atmospheric and Space Physics (LASP) studies solar, atmospheric, planetary, and space sciences. We provide data to the public that are generally time series of either single or spectral bins of values and related information such as uncertainties. Some values are direct measurements and others are algorithmic models.

Similar to many research labs, LASP funding is mission and project oriented. As a result, lab datasets are generally created and managed separately. Information about LASP datasets (metadata) has historically been managed in a variety of non-generalizable methods on a per project basis.

Data and related information are currently provided to the public via various web pages serving a variety of purposes. The LASP Interactive Solar Irradiance Data Center (LISIRD[1]) (Wilson et al. 2009) is a web site to deliver several dozen solar irradiance and related data products to the public. Besides LISIRD, other pages provide information about LASP datasets, such as pages devoted to individual missions and Education and Outreach. Information is often duplicated,

[1] LISIRD, http://lasp.colorado.edu/lisird/.

creating maintenance headaches and the possibility for serving conflicting and out of date information. Machine readable metadata are not currently available.

To improve our metadata environment and also provide semantic capabilities, we sought to achieve the following goals and objectives:

- Provide consistent, current, accurate, semantically enabled metadata to our users via web pages.
- Provide the metadata in a linked open data fashion that supports dynamic browsing and faceted search.
- Provide the metadata to the public in machine readable format.
- Render and publish the metadata in various metadata schemas in order to federate with other relevant sites.
- Provide service interfaces for automated access to the repository.
- Provide user friendly interfaces for repository management.

We decided to build a semantically enabled metadata repository to hold the current, definitive version of the metadata. The repository would be a Single Source of Truth, with tools and a process around it, including metadata ownership and authoring capabilities and the ability to crosswalk to various metadata schemas. We needed to accomplish this with limited resources and little expertise in semantic technologies.

This paper describes the creation of the LASP Extended Metadata Repository (LEMR) and supporting tools that meet these objectives. To build LEMR, we leveraged VIVO (Krafft et al. 2010; Gewin 2009), an open source semantic web application that provides capabilities to create and manage ontologies and instance data. LISIRD was redesigned and is being rewritten to dynamically insert current metadata into web page content. Key to the redesign was AngularJS (Green and Seshadri 2013; Kozlowski and Darwin Peter 2013), an open source web application framework that provides useful functionality around code structure, dynamic page generation, and testing support. Use of these tools left a relatively small amount of work remaining to build our system. We report here on the combined use of those tools and what remained for us to build in order to meet our needs.

This paper is organized as follows. Section 2, Methodology, presents the methodology, including existing baseline capabilities and work in progress. This includes the architecture, leveraging VIVO, using and creating ontologies, initializing repository content, and creating a service layer to the repository. Section 2 also includes a discussion of client side interfaces to the repository, including example web pages populated with information from the repository.

Section 3, Next Steps, discusses the establishment of processes and additional tools for working with the repository, providing linked open data, and federating with other data centers, including the development of crosswalks to different metadata schemas.

Section 4 is a discussion section that includes: the future of the ontology developed, a discussion of the methodology, some security issues, and also linking with the University's VIVO Database. Section 5 concludes with a summary of the effort.

This effort was accomplished largely via part-time undergraduate labor. It is possible that other projects with limited resources could achieve similar capabilities with their metadata. For projects or centers with existing metadata stores, the architecture described here could be laid over those stores to achieve similar semantic capabilities.

Methodology

Figure 1 gives an overview of the LEMR architecture. Information is entered into the database via the VIVO interface and also other web interfaces and automated ingest scripts. A Fuseki[2] endpoint provides SPARQL[3] read access to database contents. Stakeholders in the project include staff that administer and maintain the repository, a possible curator of the metadata, and various scientists who 'own' a dataset's metadata and take responsibility for its upkeep.

In the following sections, we will discuss the tools we leveraged and what remained to be developed to create the repository.

Leveraging VIVO to manage ontologies and create a triplestore

VIVO is an open source semantic web application originally developed to enable the discovery of research and scholarly relationships across disciplines within an academic institution. Funded by the National Institutes of Health, the VIVO web application supports browsing and faceted[4] search capabilities of the institution's data about faculty, research areas, publications, and grants. In addition, applications like VIVO Search[5]

[2] Fuseki is a server that provides the SPARQL protocol over HTTP. See http://jena.apache.org/documentation/serving_data/.

[3] SPARQL (SPARQL), a recursive acronym for 'SPARQL Protocol and RDF Query Language', is a query language for databases that allows direct and rapid querying of a semantic database, and here independent of the VIVO interface. See SPARQL, http://en.wikipedia.org/wiki/SPARQL.

[4] VIVO provides 'facets' based on types of concepts, i.e. a 'Person' would have the facets 'Faculty Member', or 'Student', with 'Student' having further facets 'Undergraduate Student' or 'Graduate Student', etc. Users can drill down through more detailed results based on the selected sub-types / facets. See http://en.wikipedia.org/wiki/Faceted_search for more information on faceted search.

[5] See https://wiki.duraspace.org/display/VIVOSearch/VIVO+Multisite+Search.

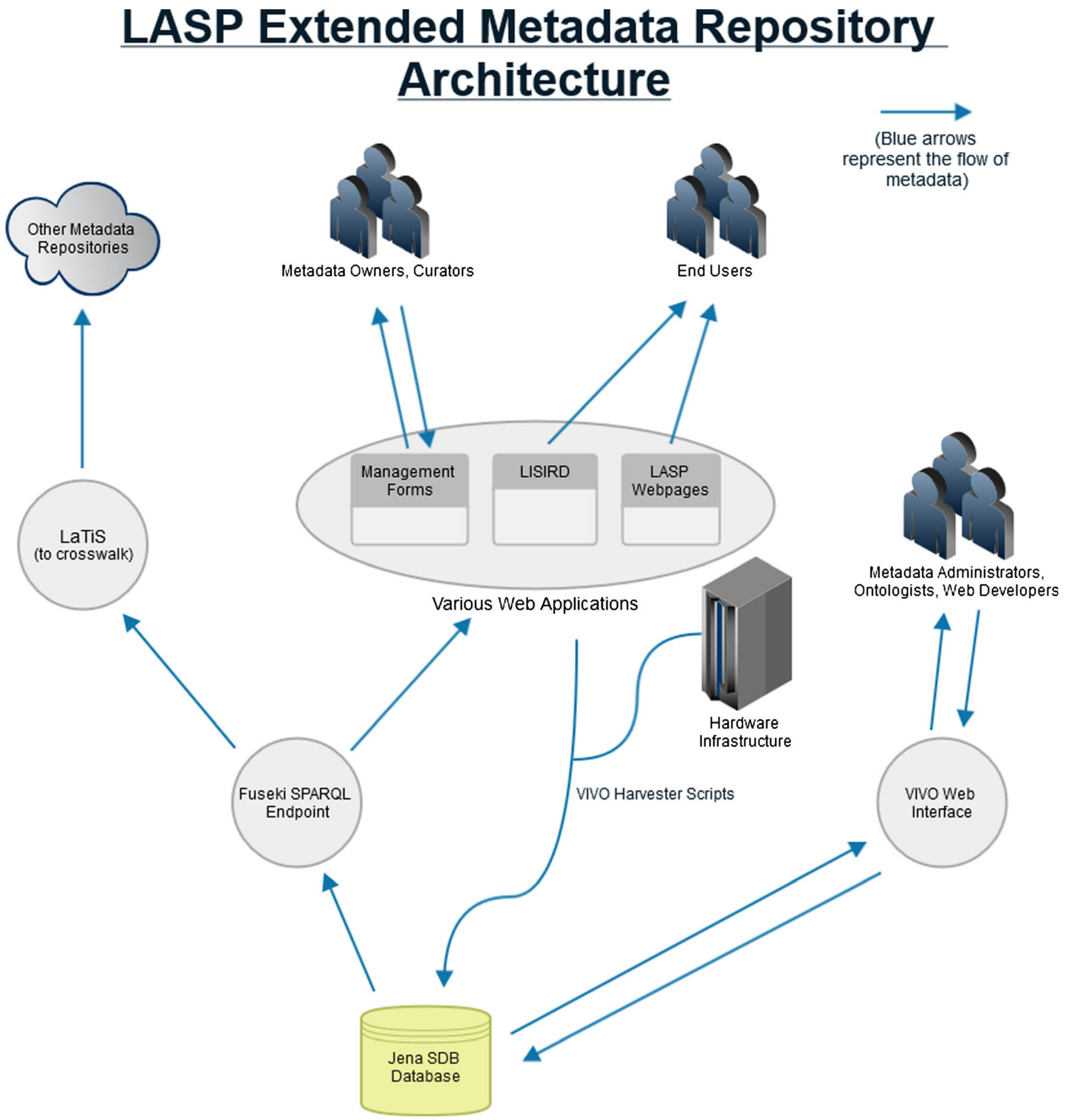

LASP Extended Metadata Repository Architecture
(Blue arrows represent the flow of metadata)
Other Metadata Repositories
Metadata Owners, Curators
End Users
LaTiS (to crosswalk)
Management Forms
LISIRD
LASP Webpages
Various Web Applications
Metadata Administrators, Ontologists, Web Developers
Hardware Infrastructure
Fuseki SPARQL Endpoint
VIVO Harvester Scripts
VIVO Web Interface
Jena SDB Database

Fig. 1 LEMR architectural diagram

allow for multi-institutional search of the data held by various institutions running VIVO.

While originally intended for scholarly research, the underlying VIVO framework is more general and broadly useful for creation and management of any semantic relationships and instances. VIVO creates and manages a database (a triplestore[6]) of semantic information, which can serve its contents in RDF[7] format (Lassila and Swick 1998). It also supports the general creation and management of ontologies,[8] as well as custom crafting of ontologies for specific needs (see The dataset ontology for an example). In addition, VIVO provides create, read, update, and deletion (or 'CRUD') capabilities for metadata record instances.

We used VIVO to create and manage our dataset–related ontologies and create instances to populate the database, a process described further below. The VIVO framework's separation of concerns allows leveraging pieces of the architecture as needed. For example, the database was built with VIVO, but we use other tools to access and manage its contents.

[6] We currently implement the default VIVO database, SDB, an Apache version of an RDF triplestore based on a MySQL database. For more information on semantic triplestores, see http://en.wikipedia.org/wiki/Triplestore.

[7] The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. For more information, see http://en.wikipedia.org/wiki/Resource_Description_Framework.

[8] In computer and information science, an ontology is a machine encoding of terms, concepts, and the relations among them. See http://semanticweb.org/wiki/Ontology for more information.

LEMR specific development tasks

While VIVO provided important functionality, some aspects of our repository had to be crafted for our purposes. In particular, it was necessary to 1) add dataset and catalog ontologies to the database and incorporate concepts from our space physics domain into them, 2) populate the repository with content, and 3) create a separate service interface to the repository to access the content independent of VIVO.

The dataset ontology

VIVO provides support for ontology management, including adding and extending ontologies and knitting them together. The VIVO platform is packaged with a number of ontologies including FOAF,[9] Bibliontology,[10] SKOS,[11] and VCard.[12] All of these ontologies are intended to capture 'person' and 'publication' metadata (such as authorship and contact information, etc.).

While these ontologies are used for some aspects of LEMR, not surprisingly, they do not provide concepts necessary for LEMR, such as missions, spacecraft, instruments, detectors and their relationships to each other, as well as to people. In an effort to leverage existing ontologies, we searched for ontologies that dealt with spacecraft and instrument metadata. The two most appropriate of these ontologies were 1) Semantic Web for Earth and Environmental Technology (SWEET) Ontology (Raskin and Pan 2005), and 2) the Virtual Solar Terrestrial Observatory (VSTO) Ontology (McGuinness et al. 2007, Fox et al. 2009).

The SWEET ontology offers a comprehensive vocabulary for a broad, high-level concept space. SWEET starts with nine top-level concepts such as 'Realm', 'Phenomena', 'Process', and 'Matter'. Each of these high-level concepts is broken down into various more specific ontologies. SWEET "is highly modular with 6000 concepts in 200 separate ontologies".[13] However, it was difficult to sift through these 6000 concepts to find terms applicable to our specific needs. Though SWEET includes concepts like 'Heliosphere', 'Magnetosphere', and 'Research', there were no terms directly relatable to a satellite, its instruments, or missions.

The VSTO ontology contains definitions of terms relevant to our space science domain. We implemented the VSTO ontology to help classify spacecraft, deployments, instruments, and detectors. However, while those concepts met our needs, the VSTO dataset and parameter extension did not serve to describe our datasets. The VSTO ontology implies that all data products are created from detectors on instruments, but LASP also offers datasets that are created algorithmically, such as computational models. In addition, VSTO distinguishes between 'data sets' and 'data products' with a number of properties related to each. The definition of those terms is often debated due to their ambiguity. We preferred to avoid them, as managing that ambiguity would introduce complications into our ontology.

Leveraging VIVO capabilities, we chose to build our own ontology better suited to our needs: an ontology general enough to eventually apply to all datasets across the lab, but also detailed enough to satisfy queries that are relevant to specific types of datasets. One use case was to be able to search for datasets covering a particular spectral range. Thus, our ontology needed to know about spectral datasets and their wavelength coverages.

Believing that it would not be possible to design an ontology today that will fit all our needs in the future, we needed a repository that could evolve. Modifying an ontology is often complicated (Klein and Fensel 2001). Extending an ontology is generally easier. A newly emerging ontology should start by describing its domain in the most generic way possible (Raskin and Pan 2005). Then, it should be easy to add further details to the generic concepts already present to accommodate the new information.

With this in mind, we designed a small, general dataset ontology that fits our needs now, with the expectation that when more specific concepts are needed in the future, they can be added relatively easily. A diagram of the initial dataset ontology is shown in Fig. 2.

Consider the DCAT[14] portion of the ontology, shown in yellow. DCAT, a W3C standard, is a small, specialized ontology designed for catalogs and datasets. Because of the simplicity and generality of DCAT, it was easy to extend it to provide additional concepts required by LEMR. For instance, each catalog needed an identifier, thus, a laspds:shortname property was added for identification on a lab-wide scope. In addition, we wanted to support a hierarchy of catalogs. At the highest level would be the catalog 'All LASP Datasets' with nested, more specific catalogs referenced within it. E.g., 'Solar Irradiance Datasets', would contain both 'Spectral Solar Irradiance Datasets', and 'Total Solar Irradiance Datasets'. To handle this, the laspds:hasSubCatalog property was added, which enabled the nesting of DCAT catalogs. These and other added properties of the LASPDS ontology are shown in Fig. 2, in green. Note that any query that would work on a repository containing "normal" DCAT datasets or catalogs would still work against our repository, as the DCAT ontology per se was not changed.

The semantic web is powerful because of reasoning or inferencing tools that 'discover' relationships between ontological concepts that are not explicitly stated in ontology. For

[9] FOAF, http://xmlns.com/foaf/spec/.

[10] Bibliontology, http://bibliontology.com/.

[11] SKOS, http://www.w3.org/2009/08/skos-reference/skos.html.

[12] VCard, http://www.w3.org/TR/vcard-rdf/.

[13] See the SWEET ontology homepage, http://sweet.jpl.nasa.gov/.

[14] DCAT, http://www.w3.org/TR/vocab-dcat/.

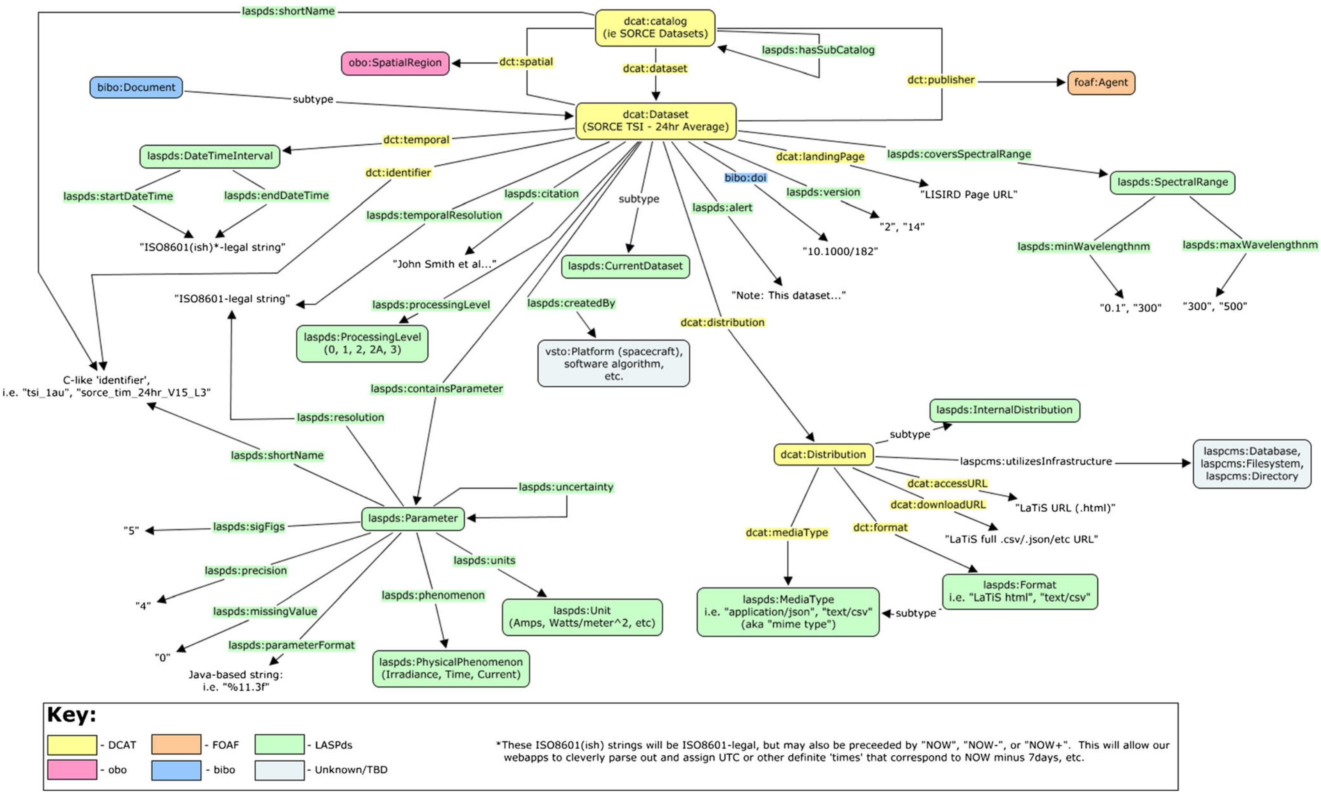


Fig. 2 LASP's dataset ontology. Note the incorporation and extension of other already established ontologies such as DCAT, FOAF, and bibliontology

LISIRD mission pages we sought to provide information about the physical quantities measured from a spacecraft. While not explicit in the ontology, this information is available via the dcat:Dataset concept, a key concept that connects the vsto:Platform (a spacecraft) and the laspds:PhysicalPhenomenon concepts via a chain of links. A more sophisticated reasoning engine could use richer mechanisms for harvesting new information or reasoning about the repository contents.

Bootstrapping the repository content

As there was no existing metadata database, it was necessary to harvest and ingest metadata by hand. Using the VIVO interface, students populated metadata fields with values, working from web pages and other sources of data documentation. Since the number of datasets and concepts is relatively small, this task was not onerous. (Before going live with the content, scientists will be asked to verify the information, described further below in section 3, 'Next Steps'.)

VIVO's harvesting tool, the VIVO Harvester,[15] has also been used to populate related information about project resource usage (an independent, internal application of LEMR). The Harvester supports bulk ingest into the triplestore. Scripts are used to gather information from various sources in the lab. That information is provided in CSV format to the Harvester, which ingests the information into LEMR.

The metadata repository service layer

The VIVO web application provides an interface for the entire semantic database. While suitable for a semantic database administrator, this complex interface is not suitable for general maintainers of repository content. For example, users whose role it is to maintain some metadata records should not be exposed to underlying technological issues having to do with ontology management and usage.

To support metadata management we needed a metadata editor application to allow domain knowledge experts to view and edit the metadata in a controlled way, without requiring a comprehensive understanding of the semantic relations imposed by the ontology. We also needed to support web pages that fetch current metadata values in creating content as those pages are constructed.

The repository needed a service inventory to support these applications. We used Fuseki, which serves RDF triplestore data via HTTP, to create a SPARQL endpoint capable of serving our semantic data. Via the SPARQL endpoint, clients can execute queries in the context of applications that limit access to only a specific subset of the metadata in the repository, such as metadata for a

[15] VIVO Harvester, https://wiki.duraspace.org/display/VIVO/VIVO+Harvester.

particular dataset that can be edited by a specific owner. Via client interfaces to this layer, both reading and writing can be controlled and limited to only the relevant portions of the database.[16]

Client side interfaces

An important use case for LEMR metadata is to provide end user access to current metadata through the LISIRD web application. This section provides further details and examples of LISIRD's use of the repository.

We identified these requirements for LISIRD:

- Read current metadata from the repository to add to web page content.
- Support metadata browsing or searching via user inputs and queries.
- As much as possible, generate web page content dynamically rather than statically.

As mentioned, queries to the database are done via the SPARQL endpoint. We use the AngularJS JavaScript framework to retrieve data from the SPARQL endpoint and build LISIRD web pages. AngularJS provides 'controllers', which contain workflow logic of the application.[17] These controllers generate dynamic SPARQL queries to gather the desired metadata for each page as the page is rendered. The corresponding metadata fields and values are returned and parsed as a JSON[18] object. AngularJS templates define the common page structure and include space for dynamic information. The templates and controllers provide a clean, manageable way to generate dynamic requests based on information provided in the URL and render web pages that incorporate information returned from the metadata repository. This way current information (such as metadata updates and important notifications like warnings) will be available whenever the page is rendered. An example of dynamically rendered metadata can be seen in Fig. 3, where a 'www.example.edu/lisird/missions/sorce' URL is passed to the AngularJS controller, which then generates a dynamic page for the SORCE mission with current information pulled in real time from the repository. (See the Appendix for the SPARQL query generated and some sample results).

With these capabilities, web pages or forms for the specialized browsing of selected subsets of the metadata can be generated dynamically. For instance, Fig. 4 demonstrates an AngularJS application to create a dynamic catalog-browsing page. This page shows hierarchically nested catalogs based on the types of datasets that LISIRD offers, currently: solar spectral irradiance, total solar irradiance, and solar indices.

Additionally, we can create pages that allow users to directly interact with the metadata, as seen in Fig. 5. That page allows users to find datasets that fall within the chosen spectral and temporal ranges. AngularJS watches for user input in the search boxes and updates the list of datasets instantly based on the input.[19]

Next steps

Processes and tools for robust metadata management

A critical aspect of LEMR usage is establishing and using good practices around it. Doing so has both technical and social aspects.

On the technical side, write access will be limited to appropriate persons via incorporation of authentication and authorization functionality in the services. These user roles will be supported:

- Metadata owner: an expert about the data that is responsible for ensuring that metadata is current and accurate. A metadata record may have multiple owners.
- Metadata curator: responsible for the repository as a whole, ensures logical consistency, etc. The repository may have multiple curators.
- Repository administrator: responsible for managing ontologies, the database, infrastructure, etc.

Formally storing metadata allows the records to be version controlled. We will investigate how to best version control the repository contents and ontologies (Klein and Fensel 2001).

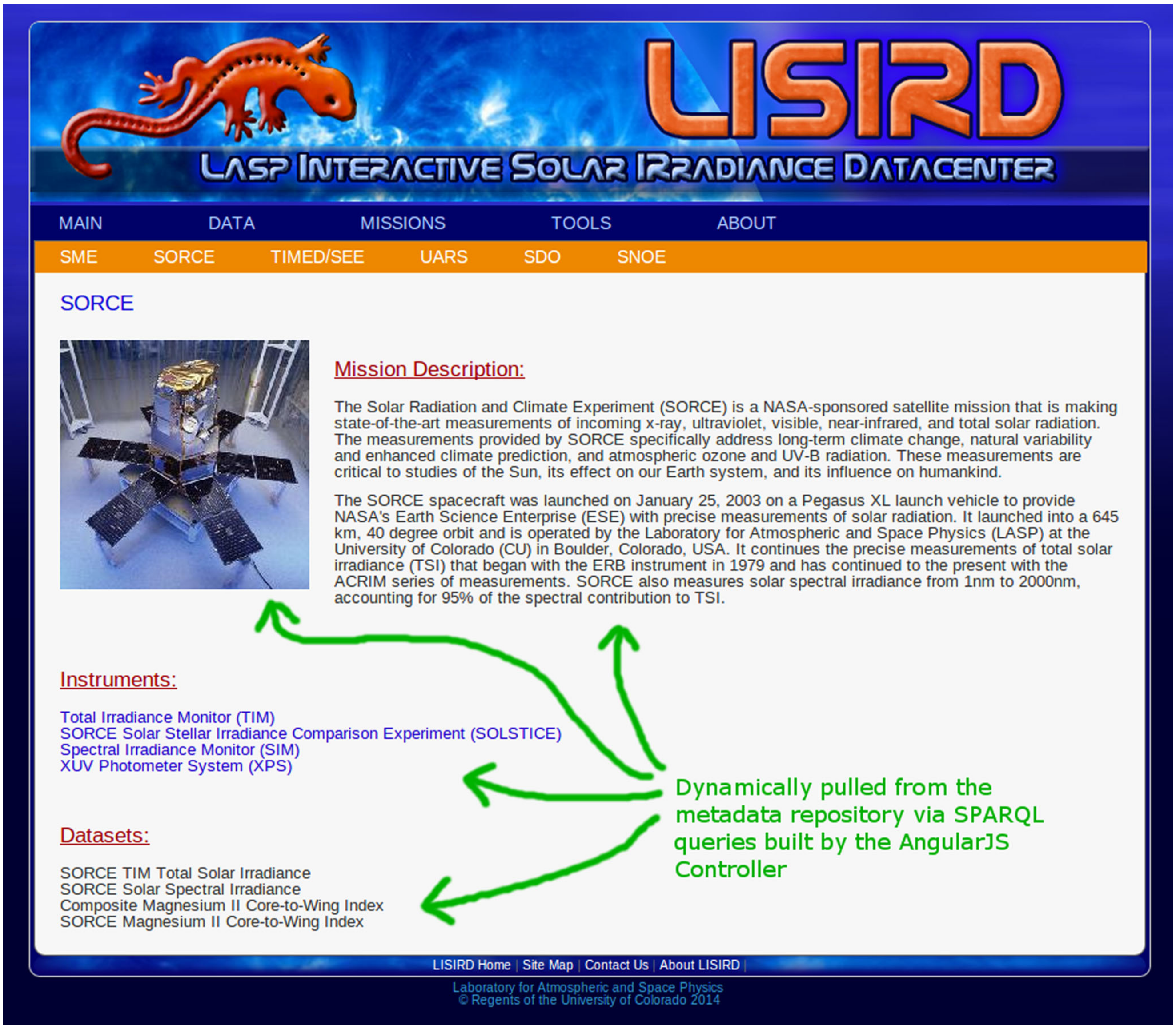
On the social side, the repository will be only as successful as its content is complete and correct. Metadata owners play a critical role in maintaining accurate information. A strong incentive for participation is making lives easier by simplifying the task of managing metadata. (The threat of going live with the information is also a powerful incentive for metadata owners to ensure accuracy). Metadata is typically managed in an ad hoc, error

[16] New versions of VIVO are planned for releases that have increased functionality such as a built-in SPARQL endpoints and a SPARQL-based update API. We will evaluate and adopt these capabilities as appropriate.

[17] AngularJS controllers model the 'C' in the MVC, or Model, View, Controller software design pattern.

[18] JSON is a lightweight data interchange format. See http://json.org.

[19] From a software development perspective, AngularJS is powerful also due to complementary tools for creating project skeletons (Yeoman) and executing both unit and end-to-end testing, such as Grunt (Grunt) and Karma (Karma), and managing dependencies (Bower). There are many on line resources, such as, http://www.sitepoint.com/kickstart-your-angularjs-development-with-yeoman-grunt-and-bower/.



LISIRD
LASP INTERACTIVE SOLAR IRRADIANCE DATACENTER
MAIN DATA MISSIONS TOOLS ABOUT
SME SORCE TIMED/SEE UARS SDO SNOE

SORCE

Mission Description:

The Solar Radiation and Climate Experiment (SORCE) is a NASA-sponsored satellite mission that is making state-of-the-art measurements of incoming x-ray, ultraviolet, visible, near-infrared, and total solar radiation. The measurements provided by SORCE specifically address long-term climate change, natural variability and enhanced climate prediction, and atmospheric ozone and UV-B radiation. These measurements are critical to studies of the Sun, its effect on our Earth system, and its influence on humankind.

The SORCE spacecraft was launched on January 25, 2003 on a Pegasus XL launch vehicle to provide NASA's Earth Science Enterprise (ESE) with precise measurements of solar radiation. It launched into a 645 km, 40 degree orbit and is operated by the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado (CU) in Boulder, Colorado, USA. It continues the precise measurements of total solar irradiance (TSI) that began with the ERB instrument in 1979 and has continued to the present with the ACRIM series of measurements. SORCE also measures solar spectral irradiance from 1nm to 2000nm, accounting for 95% of the spectral contribution to TSI.

Instruments:

Total Irradiance Monitor (TIM)
SORCE Solar Stellar Irradiance Comparison Experiment (SOLSTICE)
Spectral Irradiance Monitor (SIM)
XUV Photometer System (XPS)

Datasets:

SORCE TIM Total Solar Irradiance
SORCE Solar Spectral Irradiance
Composite Magnesium II Core-to-Wing Index
SORCE Magnesium II Core-to-Wing Index

Dynamically pulled from the metadata repository via SPARQL queries built by the AngularJS Controller

LISIRD Home | Site Map | Contact Us | About LISIRD |
Laboratory for Atmospheric and Space Physics
© Regents of the University of Colorado 2014

Fig. 3 A dynamic mission page created by an AngularJS controller. This page was created by requesting the URL "http://www.example.edu/lisird/missions/sorce". the AngularJS controller looks for the token appearing after 'missions', in this case, 'sorce', and builds a dynamic SPARQL query around that. See the Appendix for the query for this data as well as the results returned from the SPARQL endpoint

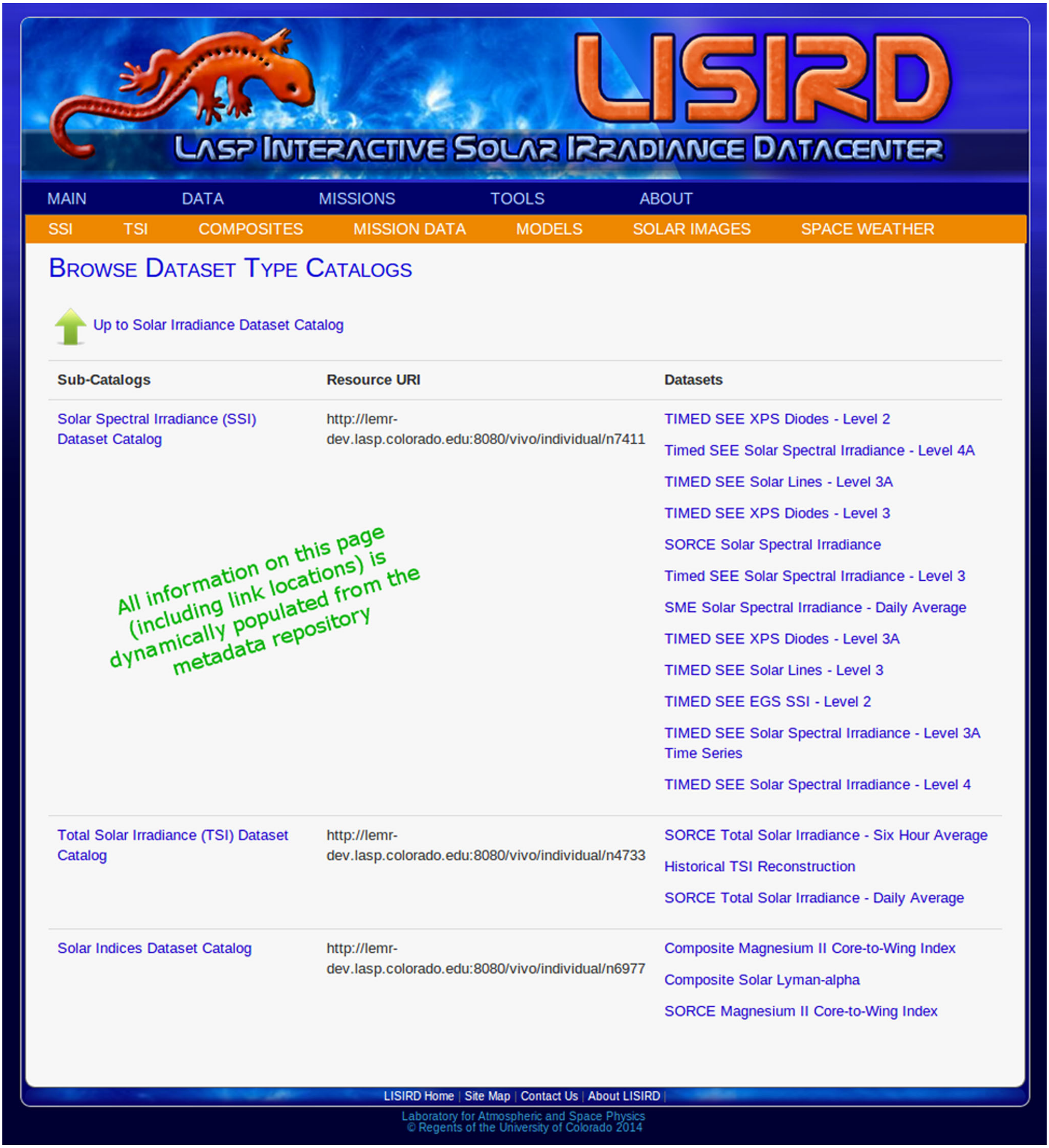
prone manner. The repository offers the alternative of storing and managing the metadata in a more formal and rigorous fashion.

We try to make the task as easy as possible for metadata owners by prepopulating fields as best we can before asking them to verify the information. A set of easy to use CRUD interfaces is under development to ease maintenance. Also, staff is available to spend time with metadata owners and help them use the forms and understand the repository. Once verified, the metadata contents are expected to change infrequently. Creation of a new metadata record can be simplified by presenting users with a mostly prepopulated form and appropriate widgets for generating values like dates. For example, a common task would be to create metadata for a new version of a dataset. A form can be presented that duplicates most information from the prior version, leaving only a few fields needing input.

Providing linked open data

To link with other semantically enabled sites, LEMR must provide linked open data. Linked open data specifications require the metadata to:

1) Be openly licensed and available publically.
2) Be linked (to itself and/or external metadata) via RDF triples.



Browse Dataset Type Catalogs

Up to Solar Irradiance Dataset Catalog

Sub-Catalogs	Resource URI	Datasets
Solar Spectral Irradiance (SSI) Dataset Catalog	http://lemr-dev.lasp.colorado.edu:8080/vivo/individual/n7411	TIMED SEE XPS Diodes - Level 2 Timed SEE Solar Spectral Irradiance - Level 4A TIMED SEE Solar Lines - Level 3A TIMED SEE XPS Diodes - Level 3 SORCE Solar Spectral Irradiance Timed SEE Solar Spectral Irradiance - Level 3 SME Solar Spectral Irradiance - Daily Average TIMED SEE XPS Diodes - Level 3A TIMED SEE Solar Lines - Level 3 TIMED SEE EGS SSI - Level 2 TIMED SEE Solar Spectral Irradiance - Level 3A Time Series TIMED SEE Solar Spectral Irradiance - Level 4
Total Solar Irradiance (TSI) Dataset Catalog	http://lemr-dev.lasp.colorado.edu:8080/vivo/individual/n4733	SORCE Total Solar Irradiance - Six Hour Average Historical TSI Reconstruction SORCE Total Solar Irradiance - Daily Average
Solar Indices Dataset Catalog	http://lemr-dev.lasp.colorado.edu:8080/vivo/individual/n6977	Composite Magnesium II Core-to-Wing Index Composite Solar Lyman-alpha SORCE Magnesium II Core-to-Wing Index

All information on this page (including link locations) is dynamically populated from the metadata repository

LISIRD Home | Site Map | Contact Us | About LISIRD |
Laboratory for Atmospheric and Space Physics
© Regents of the University of Colorado 2014

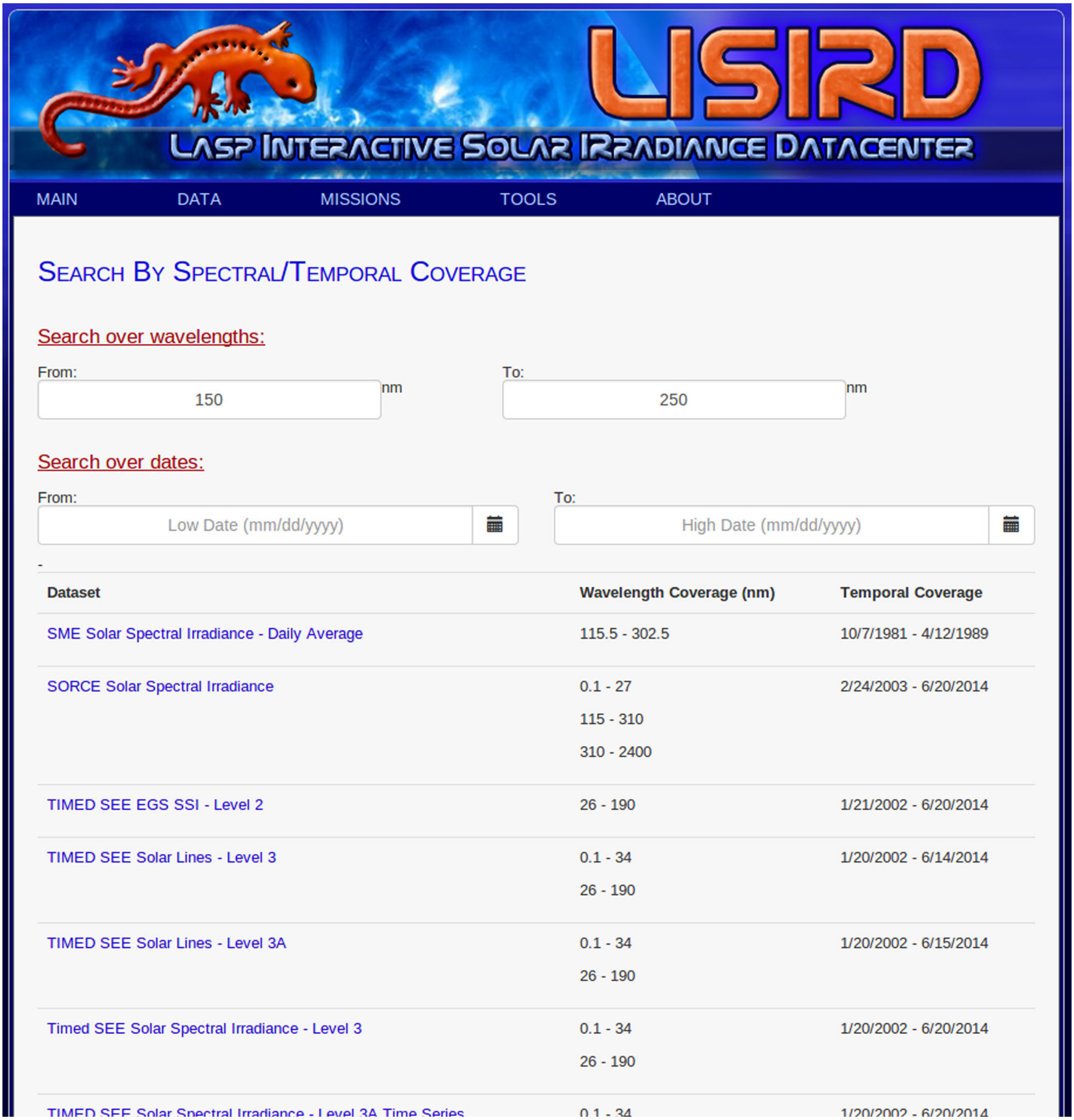
Fig. 4 A prototype of a catalog browsing page. This page was created by requesting the URL "http://www.example.edu/lisird/catalogs/types". AngularJS parses the URL and builds a SPARQL query, the results of which are used to populate cell values in the table

3) Contain resources ("things") with unique URIs.
4) Use HTTP URIs so that these resources can be referred to and looked up ("dereferenced") by people and machines.

Once our SPARQL endpoint is public, LEMR will meet these criteria.

Federating with other data centers

A major promise of the semantic web is to federate with other sites, providing a rich cross-site browsing and discovery experience. This means providing our metadata in schemas (formats) understandable to other



Search By Spectral/Temporal Coverage

Search over wavelengths:

From: 150 nm

To: 250 nm

Search over dates:

From: Low Date (mm/dd/yyyy)

To: High Date (mm/dd/yyyy)

-

Dataset	Wavelength Coverage (nm)	Temporal Coverage
SME Solar Spectral Irradiance - Daily Average	115.5 - 302.5	10/7/1981 - 4/12/1989
SORCE Solar Spectral Irradiance	0.1 - 27 115 - 310 310 - 2400	2/24/2003 - 6/20/2014
TIMED SEE EGS SSI - Level 2	26 - 190	1/21/2002 - 6/20/2014
TIMED SEE Solar Lines - Level 3	0.1 - 34 26 - 190	1/20/2002 - 6/14/2014
TIMED SEE Solar Lines - Level 3A	0.1 - 34 26 - 190	1/20/2002 - 6/15/2014
Timed SEE Solar Spectral Irradiance - Level 3	0.1 - 34 26 - 190	1/20/2002 - 6/20/2014
TIMED SEE Solar Spectral Irradiance - Level 3A Time Series	0.1 - 34	1/20/2002 - 6/20/2014

Fig. 5 An interactive search page allowing users to search for datasets that cover a specific wavelength range and/or temporal range

tools. In heliophysics that schema is SPASE (Narock and King 2008; Merka et al. 2008a; Narock et al. 2009). More generally and internationally that schema is ISO 19115.[20] Therefore, we will export the information stored in our database into ISO, SPASE, and perhaps other representations.

We will write methods to format LEMR metadata into SPASE, ISO, and other schemas and make those records publicly accessible. Subsequent registration with other entities (such as data.gov and the official SPASE registry) will allow them to harvest our metadata.

One way to build these crosswalks is to export or pull information from the repository in XML format and write XSLT transformations to the desired schemas. The SPARQL endpoint and VIVO itself can return metadata in XML,

[20] ISO 19115, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798.

though it is a flat list of triples with little ontological semantics. XSLT could be used to transform this output into other standard forms, however in some cases the mapping could be complex.

LaTiS for metadata crosswalking

Our metadata repository software stack provides another option for accessing and transforming metadata. LaTiS[21] is a data access framework of Reader and Writer plugins that leverages an intermediate 'common' representation to read data in its native format and provide data in alternative formats requested by clients. This flexibility makes it highly configurable and able to serve a large variety of datasets in a variety of formats. The LISIRD web site has for several years leveraged LaTiS as middleware to provide access to the datasets it serves. In this case LaTiS provides the data in JSON format suitable for web based interactive plots, as well as subsetting and reformatting on demand for download.

Metadata is simply another form of data servable by LaTiS. A LaTiS LEMR Reader will read dynamically from LEMR and various LaTiS Writers will provide the mappings to various other metadata schemas, including ISO and SPASE.

Discussion

Expanding the ontology

The LASPDS ontology contains a small number of concepts, just enough to support simple search capabilities over the small number of solar and related datasets currently served by LISIRD. However, LASP creates other data products in other space domains, such as the atmosphere, the magnetosphere, cosmic dust, and planetary sciences. The LASPDS ontology needs to evolve to handle queries meaningful in these subdomains.

As an ontology represents a conceptualization, it is important that it reflect a common understanding among those who would depend upon it. To expand the LASPDS ontology, we will seek input from domain experts to ensure its acceptability and utility. LASP has 65 scientists conducting research in space, a unique and valuable body of domain knowledge that we expect to leverage. Resources permitting, input from NASA's Earth Science Data Systems Work Group (ESDSWG), AGU's Earth and Space Science Informatics (ESSI) focus group, and the Earth Science Information Partnership (ESIP) will be invited.

Similar to designing software systems, designing an ontology requires careful thought and consideration of technical capabilities to be supported currently and into the future. Indeed, like software design patterns, ontology design patterns (ODPs) have been identified[22] (Hitzler 2014). To ensure maximal utility into the future, input from specialists in ontology design will be sought.

The community of space science needs better ontology choices overall in order to lower the barrier to participation in semantic technologies. In order to share the resulting ontology with the community, it will be deposited in the ESIP Ontology Portal,[23] a publically accessible repository for Earth Science ontologies. It will also be publicized within the groups mentioned above and other relevant organizations such as W3C.

Methodology discussion

In terms of functionality, LEMR provides capabilities that are a vast improvement to our prior practices. With our team of 2 students and 3 professionals, the effort has taken roughly 1.0 student time (half time while classes are in session) and .5 of an FTE for roughly one year. While having five minds involved in the design and development was very valuable, the actual labor involved was small.

Much labor went into investigating and trying existing ontologies for suitability. This may be an exercise that any organization would face. An organized library or search service of ontologies could help potential users find ontologies, but a fairly deep study comparing needed concepts and relations to those offered by an ontology is required to fully understand the matches and impedances between the two and the resulting consequences. As any domain could have an innumerable number of concepts and relations, a well-defined set of use cases is needed to constrain the problem of what concepts are needed.

VIVO's capabilities made it very easy to add a new ontology (DCAT), to integrate our concepts into it (creating the LASPDS ontology), and to link these new concepts back to the VIVO ontologies. They also provided confidence that we could recover from mistakes in ontology design and evolve the ontology. Being able to easily augment and knit ontologies is a boon to achieving functionality, validating the idea that there is value in combining 'smaller' domain ontologies for a specific solution, rather than trying to build a single large ontology that both covers a broad scope of concepts and also provides necessary specifics.

[21] LaTiS, https://github.com/dlindhol/LaTiS/wiki.

[22] The NeOn project has identified six categories of ODPs: Structural, Correspondence, Content, Reasoning, Presentation, and Lexico-Syntactic. The NeOn Project aims to "advance the state of the art in using ontologies for large-scale semantic applications". See http://ontologydesignpatterns.org/wiki/Main_Page for information about ODPs, and http://www.neon-project.org/nw/Welcome_to_the_NeOn_Project for information about NeOn.

[23] http://esipportal.cse.sc.edu:8084/ontologies.

We continue to develop our LEMR tool suite and applications using AngularJS, which provides a big lift in developing robust, dynamic applications on top of LEMR.

Security

In addition to metadata for scientific datasets, LEMR also contains related information used for LASP internal purposes. In order to selectively make portions of its content available to the public, it is necessary to extract and duplicate that content into a read-only, publically accessible location. This is expected to be straightforward, using SPARQL queries to extract the public information and a bulk copy process to dump information into the public database.

Some SPARQL queries are very CPU intensive. Making the SPARQL endpoint publicly available raises the possibility of an intentional or unintentional denial of service attack. If this becomes a problem additional solutions may need to be installed, such as classic rate limiting and access control list capabilities. However, budget constraints may limit our capability to provide this service with such protections. Keeping the service alive could depend on the extent of its use and the perceived value it provides.

Linking LEMR with the university of Colorado's VIVO database

A vision of eScience is to make datasets as accessible and useable as possible in order to enable new kinds of science, including cross-disciplinary and citizen science.

Generally dataset distribution requirements apply only to the domain of the specific research field. This leads to silos of data inaccessible to non-experts in the field, and limits the utility of the data to a small group of select users. This 'data confinement' is a barrier to interdisciplinary and novel research.

LASP is collaborating with its parent institution, the University of Colorado, to eventually link LEMR with the University's VIVO repository of researchers, publications, and grants. Such a link would provide yet more avenues of access into LASP dataset metadata via various new dimensions such as co-authorship maps or research interest concepts. A bi-directional link between our institutions will allow for co-authorship resolution in cases where the authors are not in the same department or institution. For example, if a LASP dataset resulted in a publication with co-authors in different departments, LASP would not need to generate resources and URLs for those non LASP authors, but could simply link to them in the University's database. The University is encouraging faculty and professionals to get an ORCID[24] identifier, a persistent digital identifier to identify researchers. This would likely be the concept we would use to link our two repositories.

Conclusion

With LEMR, we formalized and brought process to our metadata management and can bring current metadata to our users in an accessible way. We can now:

- Manage our metadata in a central repository with a manageable process.
- Populate web content dynamically, consistently, and accurately.
- Semantically link dataset metadata with other metadata within the lab.
- Provide browsing and faceted search of our metadata.
- Support the evolution of LEMR by evolving ontologies and their properties.

Additionally, in the near future we expect to be able to:

- Provide our metadata to the public as true linked open data.
- Semantically link dataset metadata with other linked open data from outside of the lab.
- Make our metadata available to the SPASE community, data.gov, and other communities.

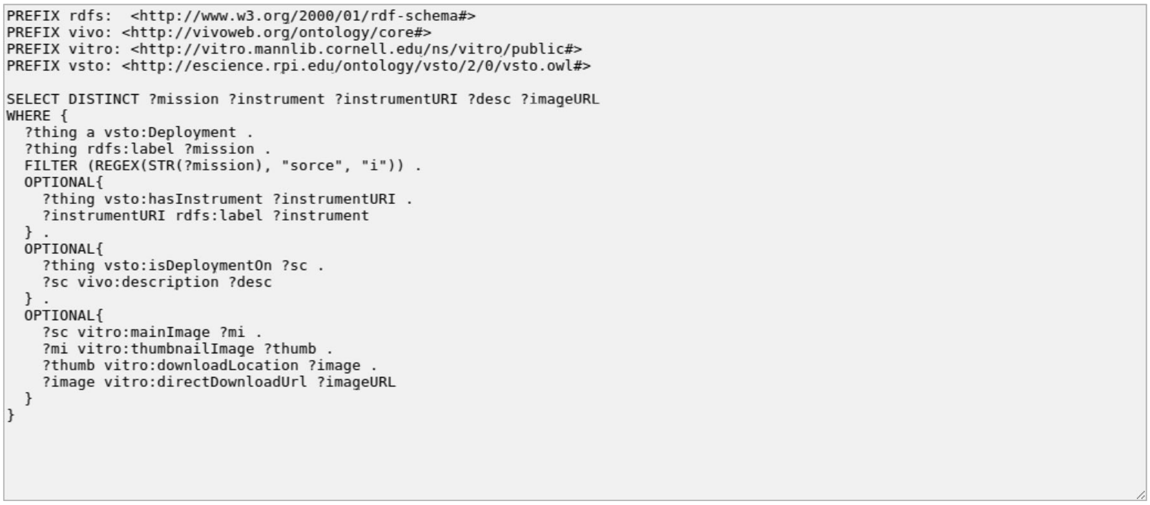
We believe access to current, definitive, verified metadata will help the users of our data understand their meaning and use them properly. Our new browsing and faceted search capabilities will help users find and understand the datasets we serve. Providing the metadata in a variety of machine-readable formats supports automated metadata access and federation. These capabilities further empower the public in finding, understanding, and using LASP datasets. It is exciting that providing these semantic capabilities is becoming within reach of the general software developer by virtue of valuable open source tools, as semantic capabilities are a key aspect of our new world of eScience.

Acknowledgments This work was supported with contributions from the NASA projects: Solar Radiation and Climate Experiment (SORCE), Multi-Satellite Ultraviolet Solar Spectral Irradiance Composite (MUSSIC), Total Solar Irradiance Center (TSIS), and Magnetospheric Multiscale (MMS).

[24] ORCID, http://orcid.org/.

Appendix

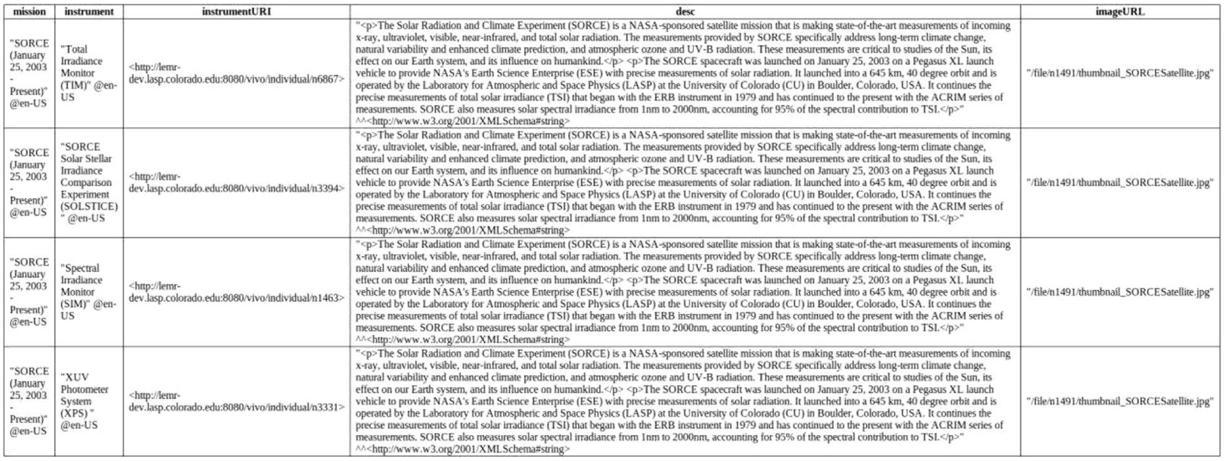
A. Example SPARQL query generated for grabbing mission metadata (datasets omitted for simplicity):



```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vivo: <http://vivoweb.org/ontology/core#>
PREFIX vitro: <http://vitro.mannlib.cornell.edu/ns/vitro/public#>
PREFIX vsto: <http://escience.rpi.edu/ontology/vsto/2/0/vsto.owl#>

SELECT DISTINCT ?mission ?instrument ?instrumentURI ?desc ?imageURL
WHERE {
  ?thing a vsto:Deployment .
  ?thing rdfs:label ?mission .
  FILTER (REGEX(STR(?mission), "sorce", "i")) .
  OPTIONAL{
    ?thing vsto:hasInstrument ?instrumentURI .
    ?instrumentURI rdfs:label ?instrument
  } .
  OPTIONAL{
    ?thing vsto:isDeploymentOn ?sc .
    ?sc vivo:description ?desc
  } .
  OPTIONAL{
    ?sc vitro:mainImage ?mi .
    ?mi vitro:thumbnailImage ?thumb .
    ?thumb vitro:downloadLocation ?image .
    ?image vitro:directDownloadUrl ?imageURL
  }
}
```

The results of the above query, which can be returned in multiple formats including JSON:



mission	instrument	instrumentURI	desc	imageURL
"SORCE (January 25, 2003 - Present)" @en-US	"Total Irradiance Monitor (TIM)" @en-US	<http://lemr-dev.lasp.colorado.edu:8080/vivo/individual/n6867>	"<p>The Solar Radiation and Climate Experiment (SORCE) is a NASA-sponsored satellite mission that is making state-of-the-art measurements of incoming x-ray, ultraviolet, visible, near-infrared, and total solar radiation. The measurements provided by SORCE specifically address long-term climate change, natural variability and enhanced climate prediction, and atmospheric ozone and UV-B radiation. These measurements are critical to studies of the Sun, its effect on our Earth system, and its influence on humankind.</p> <p>The SORCE spacecraft was launched on January 25, 2003 on a Pegasus XL launch vehicle to provide NASA's Earth Science Enterprise (ESE) with precise measurements of solar radiation. It launched into a 645 km, 40 degree orbit and is operated by the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado (CU) in Boulder, Colorado, USA. It continues the precise measurements of total solar irradiance (TSI) that began with the ERB instrument in 1979 and has continued to the present with the ACRIM series of measurements. SORCE also measures solar spectral irradiance from 1nm to 2000nm, accounting for 95% of the spectral contribution to TSI.</p>" ^^<http://www.w3.org/2001/XMLSchema#string>	"/file/n1491/thumbnail_SORCESatellite.jpg"
"SORCE (January 25, 2003 - Present)" @en-US	"SORCE Solar Stellar Irradiance Comparison Experiment (SOLSTICE) " @en-US	<http://lemr-dev.lasp.colorado.edu:8080/vivo/individual/n3394>	"<p>The Solar Radiation and Climate Experiment (SORCE) is a NASA-sponsored satellite mission that is making state-of-the-art measurements of incoming x-ray, ultraviolet, visible, near-infrared, and total solar radiation. The measurements provided by SORCE specifically address long-term climate change, natural variability and enhanced climate prediction, and atmospheric ozone and UV-B radiation. These measurements are critical to studies of the Sun, its effect on our Earth system, and its influence on humankind.</p> <p>The SORCE spacecraft was launched on January 25, 2003 on a Pegasus XL launch vehicle to provide NASA's Earth Science Enterprise (ESE) with precise measurements of solar radiation. It launched into a 645 km, 40 degree orbit and is operated by the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado (CU) in Boulder, Colorado, USA. It continues the precise measurements of total solar irradiance (TSI) that began with the ERB instrument in 1979 and has continued to the present with the ACRIM series of measurements. SORCE also measures solar spectral irradiance from 1nm to 2000nm, accounting for 95% of the spectral contribution to TSI.</p>" ^^<http://www.w3.org/2001/XMLSchema#string>	"/file/n1491/thumbnail_SORCESatellite.jpg"
"SORCE (January 25, 2003 - Present)" @en-US	"Spectral Irradiance Monitor (SIM)" @en-US	<http://lemr-dev.lasp.colorado.edu:8080/vivo/individual/n1463>	"<p>The Solar Radiation and Climate Experiment (SORCE) is a NASA-sponsored satellite mission that is making state-of-the-art measurements of incoming x-ray, ultraviolet, visible, near-infrared, and total solar radiation. The measurements provided by SORCE specifically address long-term climate change, natural variability and enhanced climate prediction, and atmospheric ozone and UV-B radiation. These measurements are critical to studies of the Sun, its effect on our Earth system, and its influence on humankind.</p> <p>The SORCE spacecraft was launched on January 25, 2003 on a Pegasus XL launch vehicle to provide NASA's Earth Science Enterprise (ESE) with precise measurements of solar radiation. It launched into a 645 km, 40 degree orbit and is operated by the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado (CU) in Boulder, Colorado, USA. It continues the precise measurements of total solar irradiance (TSI) that began with the ERB instrument in 1979 and has continued to the present with the ACRIM series of measurements. SORCE also measures solar spectral irradiance from 1nm to 2000nm, accounting for 95% of the spectral contribution to TSI.</p>" ^^<http://www.w3.org/2001/XMLSchema#string>	"/file/n1491/thumbnail_SORCESatellite.jpg"
"SORCE (January 25, 2003 - Present)" @en-US	"XUV Photometer System (XPS) " @en-US	<http://lemr-dev.lasp.colorado.edu:8080/vivo/individual/n3331>	"<p>The Solar Radiation and Climate Experiment (SORCE) is a NASA-sponsored satellite mission that is making state-of-the-art measurements of incoming x-ray, ultraviolet, visible, near-infrared, and total solar radiation. The measurements provided by SORCE specifically address long-term climate change, natural variability and enhanced climate prediction, and atmospheric ozone and UV-B radiation. These measurements are critical to studies of the Sun, its effect on our Earth system, and its influence on humankind.</p> <p>The SORCE spacecraft was launched on January 25, 2003 on a Pegasus XL launch vehicle to provide NASA's Earth Science Enterprise (ESE) with precise measurements of solar radiation. It launched into a 645 km, 40 degree orbit and is operated by the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado (CU) in Boulder, Colorado, USA. It continues the precise measurements of total solar irradiance (TSI) that began with the ERB instrument in 1979 and has continued to the present with the ACRIM series of measurements. SORCE also measures solar spectral irradiance from 1nm to 2000nm, accounting for 95% of the spectral contribution to TSI.</p>" ^^<http://www.w3.org/2001/XMLSchema#string>	"/file/n1491/thumbnail_SORCESatellite.jpg"

References

Arko R (2014) EarthCube Building Blocks: OceanLink, Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery." Presented at Open Geospatial Consortium's Technical and Planning Committee Meeting, Crystal City, VA, March, http://www.oceanlink.org/papers/OceanLink-Arko.pdf

Fox P, McGuinness DL, Cinquini L, West P, Garcia J, Benedict J, Middleton D (2009) Ontology-supported scientific data frameworks: the virtual solar terrestrial observatory experience. Comput Geosci 35(4):724–738

Gewin V (2009) Networking in VIVO, An interdisciplinary networking site for scientists. Nature 462, 123 (4 November 2009) | 10.1038/nj7269-123a

Green Brad, Seshadri Shyam (2013) *AngularJS* (1st ed.). O'Reilly Media. p. 150. (March 22, 2013) ISBN 978–1449344856.

Hitzler P (2014) "Ontology Design Patterns for Large-Scale Data Interchange and Discovery." Conference Keynote at EKAW 2014, the 29th International Conference on Knowledge Engineering and Knowledge Management, Linkoping, Sweden, November

Klein M., Fensel D (2001) "Ontology versioning on the Semantic Web." *SWWS*. 2001

Kozlowski Pawel, Darwin Peter Bacon (2013) *Mastering Web Application Development with AngularJS* (1st ed.). Packt Publishing. p. 372. (August 23, 2013) ISBN 978–1782161820.

Krafft, D., Cappadona, A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B., (2010) VIVO: Enabling National Networking of Scientists, Web Science Conf., April 26-27, 2010, Raleigh, NC

Lassila O., Swick R. R. (1998) Resource Description Framework (RDF) model and syntax specification.

McGuinness, D.L., P. Fox, L. Cinquini, P. West, J. Garcia, J.L. Benedict, D. Middleton (2007) The virtual solar-terrestrial observatory: a deployed Semantic Web application case study for scientific research. In the proceedings of the 19th Conference on Innovative Applications of Artificial Intelligence (IAAI). Vancouver, BC, Canada, July 2007, pp. 1730–1737 and AI magazine, 29, #1, pp. 65–76

Merka J., Narock T., Szabo A. (2008) Navigating through SPASE to heliospheric and magnetospheric data. Earth Science Informatics 1 (1) (April, 2008) doi:10.1007/s12145-008-0004-5

Narock T., Fox P. (2012) From Science to e-Science to Semantic e-Science: A Heliophysics case study. J Comput Geosci, Vol 46, September, 2012, Pages 248–254, doi: 10.1016/j.cageo.2011.11.018

Narock T., King T. (2008) Developing a SPASE query language. Earth Science Informatics 1 (1). doi:10.1007/s12145-008-0007-2 (April, 2008).

Narock TW, Szabo A, Merka J (2009) Using semantics to extend the space physics data environment. Comput Geosci 35(4): 791–797

Raskin RG, Pan MJ (2005) Knowledge representation in the semantic web for earth and environmental terminology (SWEET). Comput Geosci 31(9):1119–1125

Wilson, A., Lindholm, D.M., Ware DeWolfe, A., Lindholm, C., Pankratz, C.K., Snow, D.M. Woods, T.N. (2009) LISIRD 2: Applying Standards and Open Source Software in Exploring and Serving Scientific Data, AGU 2009 IN41A-1122.