IN41B-1366

Evolving LISIRD and the LASP Time Series Server (LaTiS) to Support Data Identification, Citation, and Provenance

Anne Wilson, Douglas M. Lindholm, Alexandria Ware DeWolfe,
Terry Smith, Christopher K. Pankratz, Marty Snow, Thomas N. Woods

LASP
LABORATORY FOR ATMOSPHERIC AND SPACE PHYSICS
UNIVERSITY OF COLORADO AT BOULDER

Data and Services for Meeting the Needs of Science

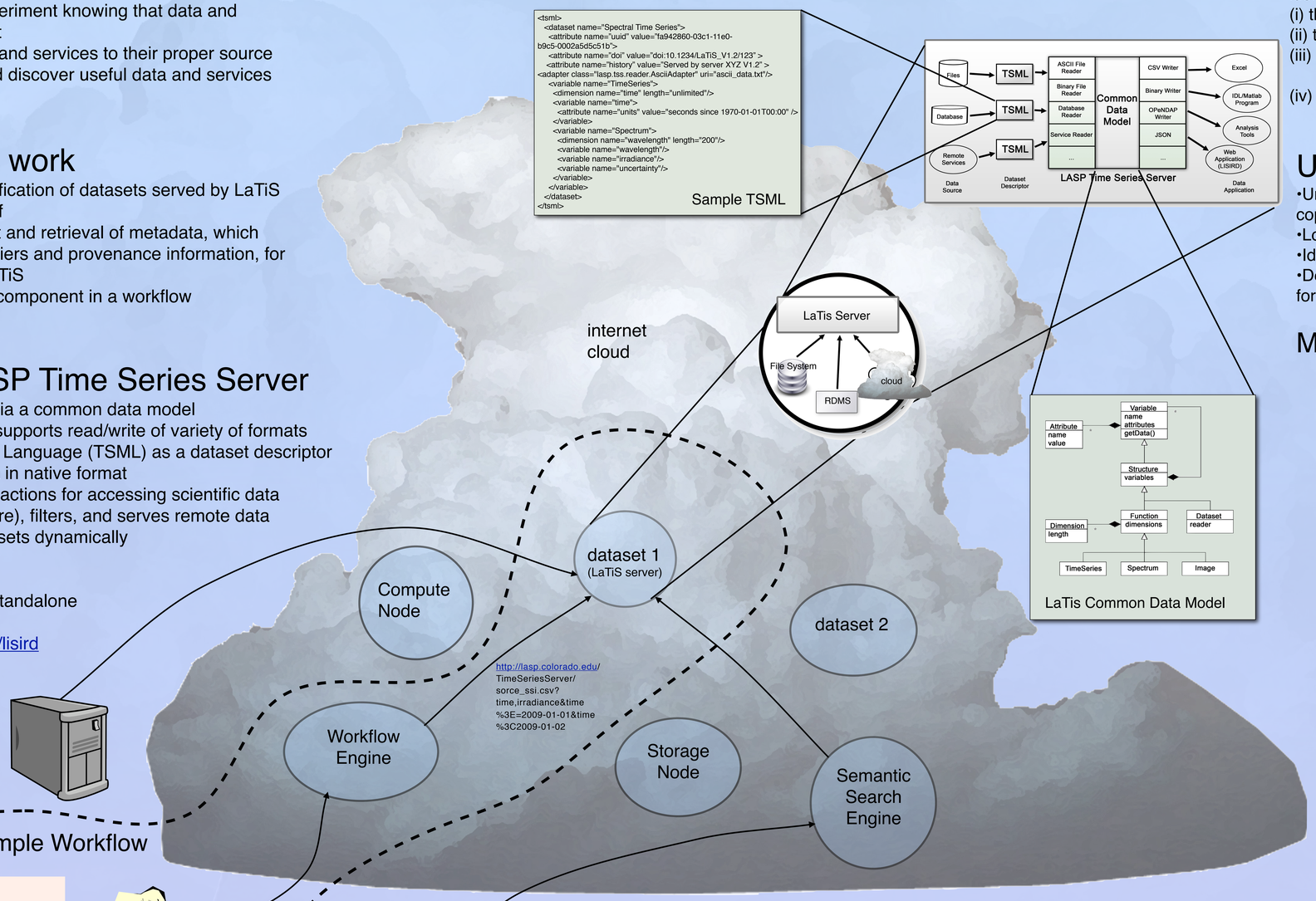
Scientists need:
- To be able to easily compose workflows to support experimentation
- To be able to repeat an experiment knowing that data and services involved are correct
- To be able to attribute data and services to their proper source
- To be able to search for and discover useful data and services

Goals of this work

- Support unique identification of datasets served by LaTiS and also of LaTiS itself
- Support management and retrieval of metadata, which includes unique identifiers and provenance information, for datasets served by LaTiS
- Consider LaTiS as a component in a workflow

LaTiS: The LASP Time Series Server

- Supports interoperability via a common data model
  - Pluggable architecture supports read/write of variety of formats
- Uses Time Series Markup Language (TSML) as a dataset descriptor
  - Enables reading of data in native format
- Provides meaningful abstractions for accessing scientific data
- Subsets, aggregates (future), filters, and serves remote data
  - Can generate new datasets dynamically
- OPeNDAP compliant
- RESTful service interface
  - Run as middleware or standalone
- Operational in LISIRD
  - http://lasp.colorado.edu/lisird



Automated Daily Updates

A Simple Workflow

Stage Data

Get dataset 1
http://lasp.colorado.edu/TimeSeriesServer/sorce_ssi.csv?time,irradiance&time%3E=2009-01-01&time%3C2009-01-02
Get dataset 2

Run Experiment → Store Results

Discovery

"solar irradiance"

References

Duerr, Ruth, 2010, "Unique Identifiers Assessment: Results", ESDSWG Annual Meeting, New Orleans, October, 2010.
Wynholds, Laura, 2010, "Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects", 6th International Digital Curation Conference, Chicago, December, 2010.

An early version of LaTiS, TSDS, is available on Sourceforge.

Acknowledgements

Unidata NetCDF-Java, including CDM: www.unidata.ucar.edu/software/netcdf-java/
OPeNDAP: www.opendap.org
Original funding for this work came from the Time Series Data Server project: tsds.net/

Conclusions

- A LaTiS request coupled with the identification of a specific LaTiS release can provide a semantically and logically concrete dataset. It also provides "an actionable mechanism for retrieval or reference" for a dataset.
- For the non self describing datasets that LaTiS serves, LaTiS must maintain dataset identifiably (in the form of separate metadata) in perpetuity.
- Cultural conventions are needed regarding how and when UIDs are assigned to datasets. This question "suggests a problem with how dataset identities are formed, such that an engagement with the definition of datasets as conceptual objects is warranted." [Wynholds, 2010]

AGU Fall 2010 Meeting

The Vision

"In order for datasets to fulfill the the roles expected of them, the following identity functions are essential for scholarly publications:
(i) the dataset is constructed as a semantically and logically concrete object,
(ii) the identity of the dataset is embedded, inherent and/or inseparable,
(iii) the identity embodies a framework of authorship, rights, and limitations, and
(iv) the identity translates into an actionable mechanism for retrieval or reference." [Wynholds, 2010]

Unique Identifiers (UIDs) Support:

- Unique and unambiguous identification of a dataset no matter which copy a user has
- Location of authoritative copy of data regardless of where currently held
- Identification of data cited in a publication
- Determination of whether two data files are the same regardless of format, "scientific equivalence" [Duerr, 2010]

Multiple identifiers may be needed [Duerr, 2010]

What Gets a UID and from Where?

"The task [of of including datasets within the scientific record] has been fraught with questions of best practice for establishing this infrastructure, especially in regards to how citations, metadata, and identifiers should be constructed." [Wynholds, 2010]

As LaTiS can create new datasets dynamically by subsetting, aggregation, and filtering, and also serve remote datasets, what is LaTiS's responsibility for UID generation? We propose these conventions:

Datasets "owned" by the data provider
- The data provider should provide a set of UIDs
- A proper subset, e.g., a subset on time or variable, of a dataset is the same dataset and should have the same UIDs
- A new version is a new dataset and thus should have new UIDs

New Datasets Generated by LaTiS
New datasets dynamically generated by LaTiS belong to the client. LaTiS can generate UIDs for them, but it is up to the client to manage them. Latis will not by default maintain metadata for every dynamically generated dataset.

Remote Datasets Served by LaTiS
LaTiS maintains metadata for serving remote datasets, which includes any UIDs provided. LaTiS will not generate UIDs for these datasets.

Uniquely Identifying LaTiS
As all software has multiple releases over time, in order to uniquely and precisely identify a dataset served by LaTiS, it is also necessary to identify which LaTiS release was used to provide the service. LaTis must uniquely identify each release.

Repeatable Dataset Generation

As a Latis request can generate a new dataset dynamically, the request for that dataset itself is a handle to the dataset. With the additional unique identification of a server, that dataset can be reconstructed. That, in turn, can be used to support experiment repeatability, dataset and service attribution, and also determination of "scientific equivalence".

Thus, a server request is a provenance artifact and should be recorded if metadata are maintained.