Web Services and Workflow Systems: Important Technological Solutions for EarthCube

Tim Ahern (IRIS, Director of Data Services),
Keith Koper (University of Utah, Chair of DMS Standing Committee),
Chad Trabant (IRIS, Director of Projects)

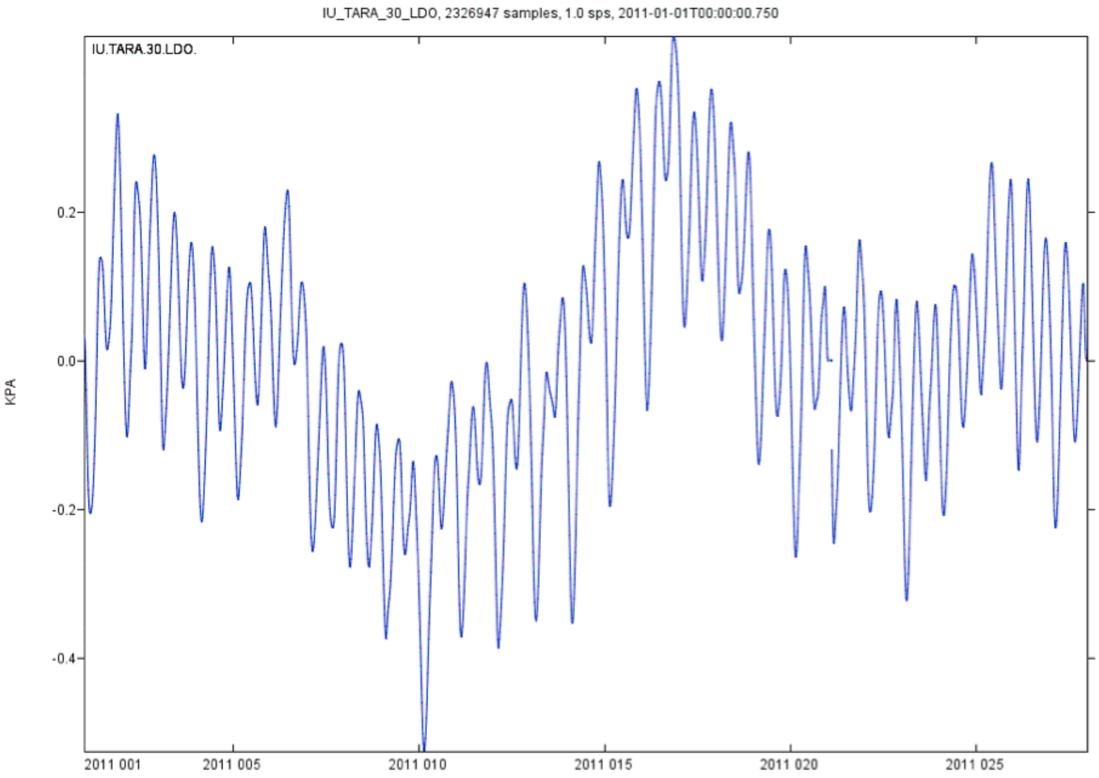
The EarthCube system will require significant flexibility to support the different needs of the geosciences community. Several organizations supported by NSF have started developing and deploying web services. Web services provide a loosely-coupled modular approach to building flexible cyberinfrastructure that could be leveraged by EarthCube.

Recently, data centers in the geosciences have begun developing service-oriented architectures (SOAs) based on web services. In many cases, such as with IRIS, one of the primary motivations for deploying SOAs was to decrease the cost and increase the efficiencies of key software components operated by these centers. Exposing selected services to the outside research community has been well received. It has been less than one year since their initial release, but web services already account for 25% (40 terabytes) of the data distributed by the IRIS DMC projected for 2011. Another major advantage of using web services is to abstract internal data formats and domain-specific proessing details, ultimately this results in increased data accessibility for a wider range of researchers.

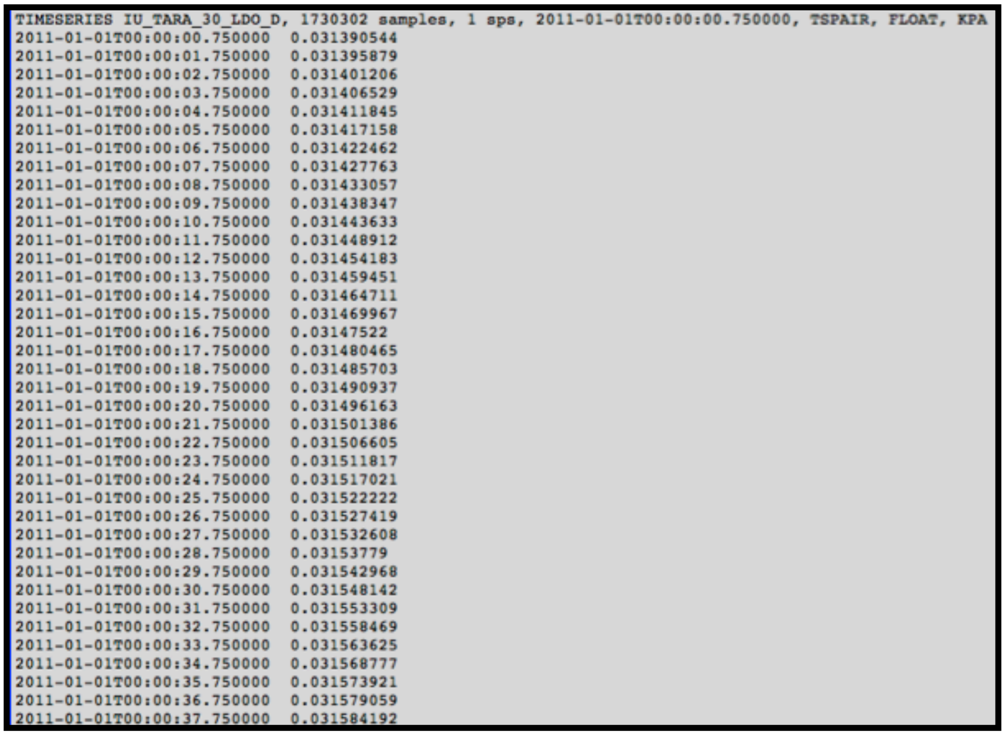
Space-Time, a unifying attribute. In general, geoscience data sets are unified in that they can be spatially referenced and have an aspect of time associated with them. Especially in data discovery, the space-time attributes of geosciences data can be heavily leveraged. A recent survey of the solid earth sciences community revealed that a vast majority (85.9%) felt that specifying latitudes and longitudes was the most important method of accessing data. 73.2% felt that specifying time was an important aspect in data discovery. Web services that take advantage of space-time attributes of geoscience data sets have already been developed at a variety of data centers. It is clear that the technologies available through web services are consistent with the important space-time attributes of geosciences data sets.

Web Services and Workflow Engines: There are two types of web service technologies in common use. These are the Simple Object Access Protocol (SOAP) web services and the Representational State Transfer (REST) web services. The simplicity of REST web services and their ease of development and use by end-users makes it compelling for their use as a base technology that should be considered for EarthCube.

Web services can be invoked as part of distributed workflows that are orchestrated by clients running locally. As part of the development activities at IRIS we have used the Trident Scientific Workflow Engine (a product of Microsoft Research) to develop workflows that connect approximately twelve web services at the DMC into useful workflows that manipulate data from the IRIS DMC. As an example, a web service that accesses raw observational time series data from the archive has been connected to a web service that applies digital signal processing algorithms that is then connected to a web service that can display the time series as a graphic. The figure below shows an example of the workflow described.



This figure shows 4 weeks of barometric pressure data from the Tarawa observing station in the SW Pacific country of Kiribati. The data have been low pass filtered, corrected for the instrument gain, and the data were then converted to a png file. While the invocation of this workflow was exceedingly simple it ultimately returned data that would be immediately useful to an atmospheric scientist, even though the data were collected by an observing station deployed primarily for use by seismologists.



```
TIMESERIES IU_TARA_30_LDO_D, 1730302 samples, 1 sps, 2011-01-01T00:00:00.750000, TSPAIR, FLOAT, KPA
2011-01-01T00:00:00.750000  0.031390544
2011-01-01T00:00:01.750000  0.031395879
2011-01-01T00:00:02.750000  0.031401206
2011-01-01T00:00:03.750000  0.031406529
2011-01-01T00:00:04.750000  0.031411845
2011-01-01T00:00:05.750000  0.031417158
2011-01-01T00:00:06.750000  0.031422462
2011-01-01T00:00:07.750000  0.031427763
2011-01-01T00:00:08.750000  0.031433057
2011-01-01T00:00:09.750000  0.031438347
2011-01-01T00:00:10.750000  0.031443633
2011-01-01T00:00:11.750000  0.031448912
2011-01-01T00:00:12.750000  0.031454183
2011-01-01T00:00:13.750000  0.031459451
2011-01-01T00:00:14.750000  0.031464711
2011-01-01T00:00:15.750000  0.031469967
2011-01-01T00:00:16.750000  0.03147522
2011-01-01T00:00:17.750000  0.031480465
2011-01-01T00:00:18.750000  0.031485703
2011-01-01T00:00:19.750000  0.031490937
2011-01-01T00:00:20.750000  0.031496163
2011-01-01T00:00:21.750000  0.031501386
2011-01-01T00:00:22.750000  0.031506605
2011-01-01T00:00:23.750000  0.031511817
2011-01-01T00:00:24.750000  0.031517021
2011-01-01T00:00:25.750000  0.031522222
2011-01-01T00:00:26.750000  0.031527419
2011-01-01T00:00:27.750000  0.031532608
2011-01-01T00:00:28.750000  0.03153779
2011-01-01T00:00:29.750000  0.031542968
2011-01-01T00:00:30.750000  0.031548142
2011-01-01T00:00:31.750000  0.031553309
2011-01-01T00:00:32.750000  0.031558469
2011-01-01T00:00:33.750000  0.031563625
2011-01-01T00:00:34.750000  0.031568777
2011-01-01T00:00:35.750000  0.031573921
2011-01-01T00:00:36.750000  0.031579059
2011-01-01T00:00:37.750000  0.031584192
```

The same data represented in the waveform plot above can be made available as an ASCII file that is readily understood to scientists outside the domain. Units of the values are in kilopascals, and the time series is easily identified as time-value pairs.

Caching. To optimize performance it is essential to minimize the amount of data that is transferred over the Internet. For this reason EarthCube should support the concept of local caching, where the artifact produced by one step in a workflow can be used as the input for the next step in the workflow at the same location. Workflows should be constructed that link as many processes as possible at a given location before passing the artifact to the next location for further processing or data integration. Web services should be constructed in a manner that can leverage local caching. The concept of an EarthCube wide caching system might also be worth considering. IRIS has implemented an initial caching system called ICAB (http://www.iris.edu/ws/icab.html) that has proven effective as a local cache used by REST web services at IRIS.

Types of General Use Web Services. Preliminary planning at IRIS and its international partners has identified the importance of a small set of web service types that should be considered being developed within domains. These include web services for

- Determining data availability
- Returning relevant metadata

- Recovery of actual time series segments of interest
- Other related domain information such as earthquake catalogs
- Access to Higher Level Products

EarthCube may find general applicability in these types of web services and coordination in implementation across domains may be advantageous to EarthCube as a whole.

Using Web Services for Horizontal Integration. An important feature of web services is they can be developed in a manner where domain specific corrections that would normally be difficult for a non-specialist to apply can be automatically applied to the data. For instance, the instrument responses that are used by seismologists are complex but the instrumentation correction can be implemented as a web service, and has been. This allows non-domain specialists to gain access to data in a form that is more ready to use.

Another example of this type of correction might be data from various labs. Many fields have noted systematic differences between measurements made at different labs. These systematic differences are often known to those within a domain but not well known outside the domain. Web services can be developed to allow corrections for these inter-laboratory differences.

Formats are often an impediment toward data sharing between domains (horizontal integration). Another example of where web services can assist in the horizontal integration of data sets across domains is in the ability for web services to deliver data in a variety of formats as requested by the client. A survey of the solid earth community for instance indicated a strong preference for sharing data across geoscience domains in easy to understand formats. In fact, the most common answer to the question "Which best describes your needs regarding formats for data outside your domain" was **"I prefer my data in an easily understood simple ASCII format".**

By offering data in multiple formats the barriers to horizontal integration are greatly reduced. The domain scientist can decide whether a simple format will suffice when accessing data across domains or if it is better to request data in the most common format within the data's domain.

By exposing data and metadata holdings in the form of web services, the barriers between research scientists and the data they need for their research are lowered. Leveraging the technology of the Internet is very consistent with making data available through simple easy-to-use web services as well. If we make data available through easy-to-use web services we believe scientists will find unique and enabling ways to combine the data in much the same way that HTML and web browsers eased access to a huge amount of information. Web services are likely to contribute to data mashups over the Internet, and EarthCube will make significant contributions to data integration across geoscience domains and the science that will result from this integration.