Web Services Concept Group Roadmap

Version as of August 20, 2012 11:00 AM Pacific Time

Executive Summary

Integration of data and information from diverse sources is becoming more important within the geosciences, as no single domain can provide the information necessary to support many research endeavors. Methods to expose information assets, no matter where they reside (e.g. individual PI systems or at various data centers), through web services lies at the heart of the Service Based Integration Platform for EarthCube (SBIP-E) concept group's ideas. We wish to remain somewhat agnostic about the specific type of web service (e.g. OGC web services, OpenDap services, SOAP based web services, or RESTful web services). We will leverage web services as they currently exist while simultaneously promoting what we term "Simple Web Services" where such services do not exist. By developing standardized methods of documentation, service discovery, standard parameter naming conventions, etc. we believe we can have a significant impact on exposing and providing access to a huge wealth of information assets. While we wish to actively support machine-to-machine interactions, we at the same time do not wish to eliminate the role of the human in this process. Experience within our group shows that significant progress can be made by simply exposing the information assets in uniform ways, providing information in understandable formats (e.g. XML, CSV, etc.), and making the interaction between data resource and scientists similar for different types of information remarkably improves the use of the data in an interdisciplinary sense.

SBIP-E proposes to first deploy web services at the various 9 partners in our group. These include the geoscience domains of marine geophysics, atmospheric sciences, geodesy, hydrology, seismology, geochemistry, bathymetry, petrology, paleobiology, sedimentary geology, geodynamics, bedrock geological maps sets, and marine underwater geophysical data. All of these data assets will be made available through web services as part of the first year's activities. Additionally higher-level information products for some of these domains will also be made accessible through web services.

After initial demonstrations between the SBIP-E partners we also wish to extend simple web services to other data providers that we engage. The development of a turnkey system that can be easily deployed to a group without a significant IT staff is a key development the SBIP-E group will pursue. Results of research in the form of products or publications will be easily exposed through such systems. Within EarthCube considerable attention is being given to the long tail of science issue that addresses data and information products that individuals produce at their own institutions and have no simple way to expose or distribute the information asset. The turnkey system addresses this issue. The existence of the RAMADDA system, initially developed by Unidata, seems to possess most of the attributes of the turnkey system SBIP-E envisions. We hope to work with the RAMADDA developers to incorporate some of the key SBIP-E concepts into the RAMADDA system.

1. Purpose

Introduction - Across all areas of research in the Geosciences there is an increasing effort to utilize integrative methods, drawing upon many different disciplines simultaneously, to improve our understanding of the Earth. One key aspect of the EarthCube project will be to allow all participants to easily share their research (in the form of data, products, tools, algorithms, etc.). We propose that a Service Based Integration Platform for EarthCube (SBIP-E) will enable community-wide commitments to sharing via standardized methods, making these methods understandable and usable in a rich array of trans-domain contexts, from discovery and workflows to visualization and publication.

This project will enhance those efforts by focusing on Web Services. We will show how a loose framework of services will allow both vertical and horizontal integration of some of the fundamental components of science, including experimental results, detailed measurements, large data sets, and computational tools. We will investigate how a loosely coupled set of Web Services (WS) can allow scientists to share these components.

By loosely coupled we mean that the remarkably diverse capabilities and technologies across EarthCube can be easily connected together. These range from simple (browser-based) functions under human control, to more complex actions that require interfacing with models, workflows, brokering systems, inference engines, even as-yet-undefined components built in a variety of programming languages. All these forms of access can be combined under a common Service Oriented Architecture (SOA). We also emphasize that many types of research groups can make available their particular expertise in the form of web services, without changes to the underlying information or techniques.

One of the most important aspects of this set of services will be to ensure harmonization of searching for, and discovery of, new information. We aim to achieve maximum utility by developing a '4-D template' for defining detailed search patterns, in both time and space, that can be applied broadly to many types of data sets. There are certain complications with respect to conventions and standards that we will address. For example, a 'polygon' is a well-known type from basic geometry, yet how to express a polygon has many variations across various computational systems like databases, GIS's, visualization programs, mathematical libraries, etc. This project will utilize community standards, formats, and components to ensure a system that is easy to understand and use.

The primary motivation of this effort is to provide new tools that can help answer specific science questions. We will show how diverse data sets can be brought together in a few test cases, or illustrations. From those detailed, usually discipline-specific, illustrations we can draw common principles. And from those specific test cases we can generalize solutions (standards, tools, frameworks, turn-key systems, etc.) that the geoscience community can use.

Communities - Every aspect of the Earth has captured our imaginations and compelled deeper understanding. We envision a system of services that allows data and tools from all geoscience sub-disciplines. We encourage scientists and researchers of all types (individuals, institutions, large data centers) to participate in this new form of science. We hope that all these communities can make use of, and contribute to, these web services.

Initially data providers with existing expertise in connecting end-users to information (centers like IRIS, NCAR, UNAVCO, etc.), can guide the coordination efforts. We will expand upon that knowledge to build a set of service methods that will allow others, especially non-experts, to share information. We hope to include tool builders with specialized algorithms or processes, including those with experimental and novel tools. In addition, individual researchers or groups with specialized data or tools will be encouraged to make those unique services available to the EarthCube community at large.

Using a few examples, we will show how geoscience researchers, especially those unfamiliar with web services, can expose their own data and tools to the larger communities. We will create a type of framework so that these groups can easily build their own web service. We propose activities like user contributed wikis, webinars, and workshops, to improve community involvement.

Technical Areas of the Roadmap - To create and develop this type of network of Web Services, many technical details will need to be set. Many options for transmitting information over the Internet exist, and many technologies are useful. This SBIP group will adopt those technologies, standards, and formats that facilitate the greatest amount of communication across geoscience research. In addition we will develop a set of search mechanisms (so called "URL Builders") to allow those unfamiliar with detailed searching patterns to quickly construct a search method to utilize the various web services. Much of the technical details will be discussed in the following sections of the Roadmap.

As a starting point, our plan asserts that EarthCube web services should be largely stateless (i.e., there are no client-server "sessions") and invoked via URLs whose forms adhere to widely agreed rules. Major milestones in this roadmap address (potential) standardization of the input arguments and the output responses associated with these web services. Clearly, EarthCube-wide agreement on these will entail intersections with Roadmaps for governance and for the various technical capabilities that employ web services as interface methodologies.

Improvements to the present state-of-the-art - For generations research in geosciences has been conducted under somewhat artificial, if useful, discipline-based divisions. To advance the state of the art of scientific research of the Earth it is clear that a more integrative approach is useful. The current situation is a group of diverse data providers with an almost equally diverse set of access methods. This project aims to unify access across disciplines. We envision a system of web services that will enable a wide variety of horizontal and vertical integration of geoscience data and tools, connecting the variety of elements together, enhancing interoperability.

For example, we imagine connections to many existing, or as yet to be created computational systems, including: primarily graphical data viewing and processing programs (GIS's, IDV, GeoMapApp, GPlates); integrated workflow engines (Kepler); pure computational systems (Matlab, R, Matematica); scripting languages (Python, Perl, PHP); and especially, "Mashups" or similar non-traditional mixings of information that derives from multiple independent sources.

Improvement to Community Capabilities - Abstract ideas and well-planned technological scenarios are essential to the scientific process, and they must advance our understanding of scientific questions to be useful. With this motivation we will develop a few key examples highlighting how a set of web services can help the EarthCube community. By focusing on a few examples with this new integrative approach to specific science questions we will show how this system can apply to other areas of geoscience research.

Using some of the existing web services from IRIS, we will show how questions in geophysics can be better answered through vertical integration, from raw measurement up, through levels of processing, to final product.
We will show how the synthesis of diverse datasets across deep time can serve to "fill in the gaps" between traditional disciplines. For example: seismic tomographic images (in the form of a 3D volume data set) can be combined with stratigraphic and geochemical data (in the form of discrete points and polygons). The depth-coded seismic tomography will be mapped to age-coded horizontal slices. These can then be used to highlight those tomographic structures that were at, or near, the surface during the time of interest. These structures can be easily correlated with volcanic rocks, and stratigraphic layers which formed at that time. All of this data can then be reconstructed through time to show the evolution of the tectonic system.

We would like to develop an example from the oceans and atmosphere communities that could take a spatial polygon and a specified time interval, and return sea surface and land surface statistical moments (e.g., mean annual temperature, variance, range etc.) that meaningfully summarize the sometimes voluminous model output and/or sensor data that are available.

2. Communication

3. Challenges

The key capability offered by web services in the context of EarthCube is facilitation of horizontal integration of a broad range of diverse data in a manner that offers a consistent experience to the user. Web services are already well represented among the user services-enabling technologies built by multiple facilities, academic departments, government labs and, research institutes within the Earth sciences. The driver for ongoing development and use of these technologies is that well-designed services have succeeded as an efficient abstraction layer that can be reused. Web services are already being leveraged within a broad range of disciplines for vertical integration and have also been exploited for horizontal integration. It is the capability of web services to facilitate vertical and horizontal integrations along with the increasing need for such integration within the context of EarthCube that is the strongest driver for the SBIP concept group. The large value of Earth science data sets in the 'long tail', and the potential for a web services interoperability platform to tap into these resources while requiring minimal investment of the providers is another driver.

Trends through time in this arena have included a decades-long evolution in technologies (CORBA, SOAP, OGC and REST [expand or refer to status]). While the advantages of REST and simple web services have become apparent, there will inevitably be evolution in this arena, and the exploitation of these services within EarthCube must be expected to evolve along with the technology. The challenge comes in ensuring that resources allow evolving the underlying technologies built to support EarthCube.

During the recent expansion of web services and the underlying infrastructure (databases, web protocols) within entities supporting a broad range of Earth science domains, efforts to establish best practices and standardization of methods, semantics, syntax […] have been undertaken, typically within a single domain. To a large extent, these efforts have succeeded within certain domains by tailoring the infrastructure to the unique user requirements within these domains; a side-effect of this success, and the challenge for SBIP, is that efforts to date in isolated domains have led to a divergence in methodologies from domain to domain. Even within domains, where community efforts have defined core infrastructure (database schema for example) and multiple groups start off adhering to an agreed structure, over time and in response to perceived unique needs, for example simplification of effort or maximizing productivity, services still have a tendency toward divergence of capabilities and methods.

Adoption of SBIP among users represents a significant challenge simply due to the frequently-encountered inertia involved of new approaches and the potential for users to be easily discouraged in facing the learning curve of new ways of tackling a problem; this challenge must be addressed through strong communication at the design and development of the web services abstraction layer and of the interoperability platforms built on top of the services between users and developers. During rollout of new services, managing user expectations through ongoing communication is essential.

4. Requirements

The process that we recommend for obtaining the necessary technical, conceptual, and community requirements for a service-oriented architecture component of EarthCube is threefold.

First, a comprehensive survey of existing community data holdings and currently deployed web services, with emphasis on those enabling the return of higher-level data and knowledge products (e.g., results derived from processed data), is recommended. Considerable effort has been made by some communities in developing web services, principally for their internal use, and the EarthCube effort can and should learn from these initiatives.

Second, a core list of requirements for the web service-based component of EarthCube should be identified by examining existing web services for technical and structural commonalities and by following the guidance provided by current EarthCube working groups (e.g., Brokering, Workflow, and Interoperability). We anticipate that the core technical requirements of the service-based component of EarthCube, in general terms, will involve 1) a common form of service description that is human and machine readable, 2) a common method of invocation based on space-time constraints, such that the same space-time arguments can be successfully passed to and be understood by all involved communities, and 3) a common form of output, with the possibility of output-format specification.

Third, the geoscientists and software developers who will ultimately use the web service component of EarthCube should be solicited for their needs, particularly with respect to modes of invocation and preferred output formats. It is recommended that emphasis be placed on simplicity in design and implementation and on the ability of web services to facilitate horizontal integration of geoscience data and knowledge products that are maintained by separate communities, not on solving the specific workflow and/or data access needs of any one community. It is also acknowledged that there are some geoscience communities that may not currently have a central data archiving or database facility upon which to create a service based architecture for EarthCube. The services component of EarthCube should, therefore, include a mechanism to engage these 'long tail' geoscientists who individually have data, results, and knowledge that would both benefit from and contribute to EarthCube.

Overall, we view some of the key requirements and priorities for the SBIP-E group activities to be similar to those outlined in the service-oriented architecture manifesto#:

- Science value over technical solution
- Strategic community-wide goals over project-specific benefits
- Intrinsic interoperability over custom integration
- Shared services over specific-purpose implementations
- Flexibility over optimization
- Evolutionary refinement over pursuit of initial perfection

5. Status

The current state of geo-informatics includes many examples of web services that are effective and adhere to key principles, almost as articulated in the introduction: 1) many are stateless (i.e., there are no client-server "sessions" and operations are idempotent), and 2) the services are invoked via URLs that adhere to simple sets of rules, though these rules are by no means "widely agreed" across all geoscience domains. The types of services provided may be grouped into three categories as outlined below.

End-to-End Data Analysis and Visualization - Some current web services (see, e.g., the Live Access Server from the Pacific Marine Environment Laboratory at XXXX and many others) are geared toward providing users an essentially complete experience from data discovery and analysis to displaying results as graphs, map views, and the like. Though likely to be part of EarthCube, such services do not generally function as building blocks for new capabilities, so they are not addressed in this particular roadmap.

Workflow Actors Invoked as Web Services - Most relevant to this roadmap are services performing discrete actions that may be combined into the natural workflows of scientific learning and advancement, workflows that are presently mediated largely by humans rather than computers. Functions supported by current geo-informatics web services include: dataset discovery; dataset access (including access to schemas, metadata and content as well as subsetting services); dataset aggregation and repurposing; and a variety of computational functions. For reasons of practicality, computational functions are beyond the scope of this roadmap.

Compound Web Services - Some current capabilities may be characterized as web services that act on other web services, such as URL builders (such as XXXX at IRIS) and simple workflow assemblers (see [About ifttt](), e.g.).

Domain-Specific Examples - Table 1 lists a representative sample of geoscience datasets accessible via web services, annotated to indicate sources, key characteristics and, by name, the web-service technologies by which the data may be accessed. Table 2 characterizes those access-providing services, highlighting key technologies and prominent features.

[insert tables here]

As may be discerned from these tables the DAP protocol--which may be deployed using free, open-source software from Unidata, OPeNDAP and others--is notable for its pattern of multi-domain usage. Collaborative OPeNDAP-Unidata efforts (funded by NOAA) are underway to further enhance this pattern of multi-domain utility and of protocol implementation via reusable software.

Gap Analysis - Listed below are weaknesses evident in the current state of web services, relative to the purpose of this roadmap. As previously stated, this purpose is focused on data-access services, specifically including those for dataset discovery, aggregation, access (including retrievals of schemas, metadata and content, often with subsetting) and repurposing. This list of weakness is partitioned into three groups.

- Forms of Service Description
    - By and large the listed dataset collections employ disparate forms of service description, making contrasts and comparisons difficult, especially across disciplinary domains.
    - As a general rule, descriptions of web-service capabilities are less than ideal for use in contemporary tools of semantic analysis, such as inference engines.
- Patterns of Invocation (with Arguments)
    - With the exception of the various DAP implementations, there is little in common, syntactically or semantically, across the methods used to invoke web services.
    - Contrary to the principle put forth in the introduction, URLs employed to invoke currently available services do not adhere to simple, widely-agreed rules, especially across domains.
        - Though many services allow users to choose among output types, there is no uniformity in the specification of those types.
        - Though many services allow users to impose space-time constraints, there is little uniformity in the syntax or semantics of the associated constraint expressions.
        - Vocabularies for other kinds of constraints (such as parameter names, e.g.) often differ, especially across domains.
- Forms of Output

- Though some services allow users to inspect the schemas and metadata of datasets and collections--without accessing their contents--the forms in which this information is delivered are quite variable, and they are sometimes designed more for human consumption than for automation, such as semantic inferencing.
- Few web-based data-access methods offer means to gain high-level summarizations, inventories or density maps of datasets whose contents (not just their metadata) satisfy user-specified conditions.
- Though many services offer human readable outputs, jargon and terse abbreviations (in dataset naming, for example) make dataset discovery and use difficult, especially across domains.
- Finally, there is immense variation in the forms used by web services to deliver dataset content (with or without subsetting). Perhaps the most important exceptions are two: 1) significant numbers of services package their output to allow access via netCDF APIs (in multiple languages), consistent with the Climate-Forecast (CF) conventions;
  2) in a few cases (such as DAP servers from Unidata and OPeNDAP) on-the-fly format translations are among the users' options.

It may be noted that addressing certain of these weaknesses may help address the others. For example, output forms that ease the application of inferencing services might mitigate the effects of vocabulary and related disparities. Furthermore, this roadmap is built in part on the premise that well defined and widely agreed web services, especially for trans-domain data access, will help establish a service-oriented architecture for EarthCube, with benefits that are clear from the many contemporary applications of this approach.

6. Solutions

EarthCube technologies will always evolve and therefore we can never assume that EarthCube solutions will remain static. A variety of web services have already come into common use and largely disappeared. Twenty years ago technologies like CORBA were commonly used within enterprises. CORBA was truly a web service but their Enterprise nature limited their adoption and widespread use. Perhaps ten years ago the concept of Service Oriented Architectures (SOAs) were gaining acceptance. This approach introduced the concept of contracts; no matter how the implementation may have changed, the contract guaranteeing a specific result when accessed in the same way allowed service interfaces to remain constant even as the actual implementations could change as necessary. Most SOAs are implemented through web services at the present time.

The service based integration platform (SBIP) concept group endorses all types of web services with the understanding that they may need to evolve with time, as is the case with all technologies. The SBIP team has identified a few types of existing web services with varying styles and implementations. The unifying feature of these services are that the interactions, in general, appear as a URL followed by query parameters and service requests appearing in URL strings. Specifically the SBIP team has identified three specific classes of web services that appear to have general utility. The Open Geospatial Consortium (OGC) web services are already widely used in geosciences and there is active work on defining and refining the standards and definitions. Efforts such as OPenDap are support OGC and other standards and the SBIP effort recommends supporting OPenDap within EarthCube.

Simple Web Services (SWS):
In addition to OGC and OpenDap services the SBIP team is also pursuing web services that are easily deployed at both data centers or on systems operated by individual researchers. Termed "Simple Web Services" the SBIP team would like to deploy prototypes of these simple web services across the 9 data providers within the SBIP group. By deploying simple web services we will gain an understanding of various aspects of their use as a method to expose data assets across the geosciences. Currently some data centers within the SBIP effort have prototypes, and in some cases, operational simple web services. To move forward within EarthCube we will promote the definition and use of standardized query parameters focusing first on space-time queries that can be used to discover data across multiple sources of data assets or to recover data from the same resources.

Standardization of Query Parameters:
SBIP will develop standard vocabularies for queries across most of the geosciences and also support flexibility for domain specific query parameters. We will propose standard methods of documenting the definition of query parameters starting with those in the 9 SBIP partners and allowing broader input for parameters specific to other disciplines. The key feature is to provide standard formats for the definitions of parameters as well as their use. IRiS has developed simple web services already as well as a strawman for how the web services are documented. Examples of this documentation can be found at www.iris.edu/ws where the entire set of services are enumerated as well as identifying the service at a high level, as well as identifying the formats available for output.

Standard Documentation and Methods to Build URLs: The documentation will also include more specific definitions and usage for specific web services. An example of this is for seismic stations with the ws-station service. (see http://www.iris.edu/ws/station/ for instance). These descriptions completely specify parameters that can be invoked, the complete specification of those parameters and any standards used for those parameters. Each web service description will include a utility called a URL Builder available from the service description page. Examples of URL builders for IRIS' simple web services can be found at URLs like http://www.iris.edu/ws/builder/station/ or from the service description page. A URL builder presents a GUI within a browser window supports the construction a URL showing the usage of the various parameters. For instance the station URL builder to recover metadata about the seismic station ANMO in Albuquerque, New Mexico for the specified time period from 1997 through 2011 would look like this:
http://www.iris.edu/ws/station/query?net=IU&sta=ANMO&loc=00&cha=LH?,BH*&starttime=1997-06-07&endtime=2011-06-07&level=sta . As fields in the GUI are supplied the URL dynamically changes to reflect the information entered. These URL builders are intended to show how to build URLs to invoke the underliying services, they are not

intended to actually recover large amounts of information but could be used for small, targeted queries.
A key feature of the simple web services effort is that it does not really take the human out of the problem, instead simple web services are intended to allow the discovery and transfer of data assets in a straightforward manner. With clear definitions of how information is returned scientists can gain access to the data they desire and be involved in the process of integrating information across disciplines.

WADL:
Simple web services will always be made available with a Web Applications Description Language (WADL) to help enable machine-to-machine interactions. How helpful these WADL descriptions will be needs to be evaluated and we will be able to do this as part of the early prototyping.

High Level Products:
The SBIP concept group believes that one of the most useful ways to make domain specific information available to a broader community is by supporting higher-level products. For instance structural geologists are not really interested in obtaining seismograms, but they are interested in gaining access to tomographic velocity modules that have been obtained as a result of seismologists conducting such studies. SBIP will support the discovery and distribution of higher level products developed by domain specialists.

Long Tail of Science and a Turnkey System:
The SBIP team will participate in the development of a turnkey system that can be easily deployed near the data assets of individual researchers. This system will expose selected data assets (including raw observational data, research results, computational tools and methods) as well as a small subset of metadata in a standard XML format that will allow support of space time queries as well as a first level of metadata that describes the information asset. We are aware of the RAMADDA system initially developed by Unidata and we think it is an excellent technology to address many of the needs of the envisioned turnkey system. Coordination between SBIP concept team members and the RAMADDA developer will likely prove beneficial to the overall EarthCube solutions needed. In particular, coordination in how the space-time queries are generated is likely quite fruitful.

Proof of Concept: **Raster Binning & Masking Services (RBinMasks)**

The EarthCube Web-Services Concept Group is examining ways to accelerate scientific advancement and learning through a simplified set of web-based data-access methods, emphasizing those that could be applicable across most or all geoscience disciplines. We focus here on one aspect of that broader discussion: how a simple class of *space-time queries* might be standardized for utility in multi-disciplinary contexts, including surprisingly complex cases where scientists in one domain seek data from other (less familiar) domains.
Conceptually, we see single queries as part of a larger process in which successive steps yield progressively clearer understandings about which

datasets, or subsets thereof, are applicable to a scientist's problem. This often requires decisions based on *data from multiple, remote and diffuse sources*, in which case the value of the aforementioned standardization is amplified.

Though merely identifying standard nomenclature for a space-time *point* would be somewhat helpful to a discipline-hopping scientist, we think it is important to address the case where the scientific interest pertains to a space-time *region*--which may have an arbitrary shape--and where the scientific results depend, not just on the *presence* of data, but also upon the *density* of pertinent data across the region. This would be useful, e.g., in the frequently encountered case where a property under study is considered a *function* over the space-time region, in which case the data values are function samples and the dataset, writ large, contains *sampled functions*. In such cases the sampling resolution can be a critical determinant in data applicability.

These considerations lead us to think that simple, standardized objects for representing density functions as *space-time rasters* could be of great value for data discovery, especially in cases that involve successively refined queries across multiple data sources.

A key aspect of our approach is to design objects that serve well as data-access inputs *and* outputs. Specifically, we are planning a type of query function that returns (rasterized) density maps and that accepts these same types of objects--namely, RBinMasks--as query parameters, serving in the latter case to define an (irregularly shaped) space-time constraint and simultaneously to specify the bins in which the returned sample densities are to be counted.

To stay within the available time and budget constraints, this proof of concept will be limited to an even narrower context by imposing some *additional* simplifying assumptions, as follows:

- ● While believing that the idea is applicable to higher-dimensional space-time cases, the planned demonstration will address only the 2-dimensional, latitude-longitude case.
- ● The demonstration rasters will be defined in a single geographic coordinate system, namely, the latitude-longitude coordinates typically associated with GPS.
- ● The raster format for the RBinMask demonstrations will be selected for ease of implementation and visualization. Eventual EarthCube standardization could occur around something entirely different.
- ● While thinking that important standardization might be achieved through widespread adoption of the RBinMasks concept, it will be demonstrated only within the context of a very few (extended) OPeNDAP clients and servers, developed by OPeNDAP staff.
- ● Though the concept likely can be extended to other cases, the demonstration will be limited to datasets in which the identification of sampled functions on space-time coordinates is completely straightforward. At a minimum this will include datasets that satisfy the widely used CF (Climate-Forecast) conventions.

•

• What we mean by proof of concept is to design, implement, and test--as an extension to OPeNDAP's widely used Data Access Protocol (DAP2)--a new form of data query. Departing from DAP2, whose response types are limited to schema-level metadata or detailed-level data values (with subsetting), the new form of query responds with rasters, called RBinMasks, that indicate the geographic densities of function samples satisfying certain query parameters.

•

• Here is a preliminary sketch of the new query-function design:

1. 1. Every query is applied to a single dataset, specified by its URL (on the DAP server)..

2. 2. Among the query parameters is the specification for a simple (equally spaced) latitude-longitude raster.

a. a. This specification may take the form of an actual raster (perhaps presented to the server via an HTTP POST command), where the value in each cell is an integer (between 0 and 255) encoded as an 8-bit byte. Cells having zero values represent geographic areas that servers must *exclude* from consideration when generating the response; cells having (any) non-zero value define bins in which function samples are to be counted, as defined below.

b. b. Alternatively the specification may take the form of a bounding box plus numbers of latitude and longitude subdivisions. In this case, each cells is assumed to have a value of 255 (indicating that it represents a bin).

3. 3. Query responses will be of several types, depending on the absence or presence of *other* query parameters as indicated in the following.

a. a. All responses will include a document that describes any output raster(s) and lists all of the "variables" in the dataset, grouped into three classes:

i. i. Variables that are known to represent geographic domain samples, i.e., their values represent latitude and/or longitude coordinates.

ii. ii. Variables that are known to represent sampled functions on the above domains, i.e., on the above latitude and/or longitude coordinates.

iii. iii. Variables that fall into neither of the above categories.

b. b. All variables listed as above will include the associated units of measure if known to the server.

c. c. Queries may include a variety of DAP2-style constraint expressions that limit consideration to subsets of the original dataset as specified by variable names and array indices. In these cases the response shall be as though the specified subset is all that exists.

d. d. If the query includes any predicate-style constraints (Temperature > 40, e.g.) on variables of type ii above, the server shall return one raster for each each such constraint, and the associated bin counts shall reflect the number of samples for which the predicate evaluates as true.

e. e. If the query does not indicate which variables to examine, a single raster of bin counts shall be returned for each coordinate variable listed under 3.a.i.

4. 4. As a general rule (but see 5 and 6 below) the bin counts returned by a query shall be encoded in a raster whose form is identical to that submitted as a query parameter. In particular, where input rasters have zeros, the output counts likewise will be zero. In cases where bin counts exceed 255, all counts in the raster shall be scaled such that the maximum count is 255 and the others are proportional. All scaling factors employed shall be documented in the responses.
5. 5. In cases where the server is unable to perform the requested binning, such as when the dataset includes no variables that can identified as samples over a latitude-longitude domain, the response shall be as in 2.a, returning no raster.
6. 6. In cases where a specified function appears (as seen by the server) to be sampled only on latitude or longitude but not both (such as when the data represent zonal means, e.g.), then the returned raster shall be collapsed to a single row or column as is appropriate for accumulating the associated bin counts. The outer boundaries of the returned raster shall remain identical to those of the input raster, which implies that the returned cells may be very long in one dimension and narrow in the other.
7.
8. This design should be considered a work in progress, as it may change substantially in the course of the proof-of-concept demonstration. In particular, changes may be necessary to enhance the value of RBinMasks in sequences of queries that reflect a user's process of gaining successively refined knowledge about (multiple, diverse) datasets of interest.

Web service Solutions:

SBIP-E sees its role as a group that finds and assists in making a variety of data assets available through web services. EarthCube web services will provide data that is scientifically meaningful, valid, and of known quality. The web services are intended to work closely with developments within the Interoperability and Brokering Concept Groups. We intend to build bridges that allow OGC and simple web services to interoperate and allow the exposure of information assets in a manner desired by an individual requesting the information. As such the web services group sees itself as a foundational capability that can be used by the Brokering and Interoperability concept groups to provide higher-level access capabilities across information assets. The web services group will do the hunting and gathering of data assets partly by supplying turnkey systems to allow access to large numbers of data collections. Web services can either be used directly by researchers or be used through other types of systems developed by the interoperability and group or be used to feed information from diverse sources into the EarthCube brokering system.

While web services concepts are still evolving we are sure they will play an important part in EarthCube's larger mission.

Community Input:

The SBIP team has members from a fairly broad cross section of the geosciences community. These include members from the solid earth, atmospheric, and the oceanic domains. We realize much broader input will be required to capture the full range of information assets supported by NSF GEO. We intend to establish both formal and informal mechanisms to insure the participation of the broad community into our activities. This will require coordination with the EarthCube governance group as well as other community groups and concept groups especially the Brokering Concept group and the Interoperability group.

7. Process

We recommend the establishment of structured, documented, community-based process(es) to select and evaluate standards, protocols, test data, use cases, etc. and promote interoperability and integration between elements of EarthCube. We may (fortunately) draw upon a large body of experience from well-established organizations, including both broad-based organizations such as the Open Geospatial Consortium (OGC; http://www.opengeospatial.org/ogc/process) and smaller discipline-based organizations such as SeaVoX (https://www.bodc.ac.uk/data/codes_and_formats/seavox/).

OGC is a consortium of academic, government, and industry partners that provides a global forum for the development of international standards for geospatial interoperability. OGC has a well-established development process for formal Members, as well as multiple opportunities for input from Non-Members.

- The process for Members includes Identifying a problem; Crafting a solution; Evaluating a proposed standard; Voting and publishing; and Encouraging implementation. Specific solution tracks (depending on scale and requirements) include a Request For Comment (RFC) where an individual or small team advances a candidates; an Interoperability Program (IP) where Members engage in sandbox/rapid prototyping; and a Working Group (WG) with structured discussion and document writing.
- Opportunities for input from Non-Members include Public forums moderated by staff; Online forms to submit Change Requests and/or new requirements; Domain Working Groups (invited speakers); Public document reviews and comment periods; and Alliances with other standards bodies.

The OGC process represents 20+ years of experience from hundreds of members, resulting in many successful standards now implemented in software (both commercial and open-source).

SeaVoX is an initiative established by the international marine science community, whose goal is to moderate code lists and controlled vocabularies to regulate the population of metadata. Its specific mandate is to manage vocabularies for the SeaDataNet program and ISO 19139 profile (except for platform codes), but it has steadily grown to encompass a wider body of vocabularies. The SeaVoX process includes Mailing list-based discussion to raise topics and reach consensus; Decisions following a “Week of silence” or else referral to the MarineXML Steering Group Chair; and Publication of vocabulary terms whose definitions may later be broadened, but never deleted.

We recommend that the establishment of an EarthCube process for Web Services development and deployment should 1) draw heavily on the experience of existing organizations such as OGC and SeaVoX, as described above, and 2) engage with other Community/Concept Groups (in particular Interoperability and Governance) and tailor the process to meet the particular needs of the EarthCube community.

8. Timeline

YEAR 1 ACTIVITIES

TECHNICAL ACTIVITIES

- Development of OGC web services as needed
- Development of Simple Web Services at Data Centers
- Development of Simple Web Services by set of small groups with specialized data and methods.
- Wrapping simple web services in OGC Processing Services (WPS)
- Work with Brokering and Interoperability Groups within the Concept Awards to develop common methods of discovery and access to information assets.

COMMUNITY ACTIVITIES

- Establish Discussion Forums specifically related to Web Services
- Establish web service short courses in conjunction with other national and regional meetings including AGU, UNAVCO annual meetings
- Vet (preliminary) web-services standards/specs in the governing committees, seeking feedback on any existing (multi-domain) realizations

YEAR 2 ACTIVITIES

- Based on earlier feedback, revise the standards/specs to inform the steps below
- Continue the development of simple web services within the SBIP participants
- Development of a turnkey system to enable deployment on systems managed by individual researchers
- Work with Governance Group toward development of the EC Organization
- Conduct surveys of the Geosciences to identify additional sources of data available in the community willing to expose data assets through Web Services
- International coordination with European and Asian data centers in geodesy and seismology to produce federated web services

YEAR 3 ACTIVITIES

- Seek broad feedback on early achievements in multi-domain and long-tail contexts, sufficient to assess the values of standards/specs and of associated software supports (such as turnkey systems and client-server libraries)
- With governing-committees, plan to scale up of the most effective software supports
- Begin engaging the broader international geosciences community through efforts such as COOPEUS
- Sponsor a “Get Connected EarthCube Workshop” during AGU or other national event

YEAR 4 ACTIVITIES

- Continue adding EarthCube partners connected via web services
- Establish full-function, sustainable user-support systems, sufficient to foster large-scale, long-term deployments

LONGER TERM ACTIVITIES
- Explore web-service extensions that embrace richer server-side functionality, applicable to cloud-computing contexts
- Maintain all specs and software to reflect the state of the art, without losing reliability or (reasonable levels of) backward compatability.

9. Management

Our team recognizes that the management of the SBIP-E concept activity, the governance structure established for EarthCube as a whole, including this project, and the relationship between the two, will be central to the success of both our efforts in support of EarthCube's overarching goals of creating a cyberinfrastructure that integrates diverse geosciences data in an open, transparent, interoperable, and inclusive manner.

Management of the SBIP-E project team will be led by PI Ahern of IRIS, who will work in close partnership with the co-investigators and their respective teams at each SBIP-E node, members of or liaisons from proximate EarthCube teams (e.g., Data Access, Cross-domain Interoperability Testbed, Brokering, and Workflows), and in synergistic ways with the overall EarthCube governance structure that will be established. Sub-teams of working groups will be established and tasked as appropriate for facilitating rapid progress or expediting decision making amongst a widely distributed team. To ensure that the direction and progress of SBIP-E activities are responsive to community needs and in alignment with EarthCube's broader mission, goals, and milestones, our roadmaps calls for the adoption of a flexible, nimble, and inclusive approach for making decisions and course corrections, coordination, dissemination, and community engagement.

Following this roadmap to successful conclusion, and achieving its overall purpose, will require the following relationships with governing framework for EarthCube, and where appropriate, the establishment of liaisons:

Decision Making - A key principle underlying this concept is that data from all EarthCube participants should be made accessible via standardized and uniformly documented web services. Hence EarthCube must embody decision-making mechanisms sufficient to instantiate such principles. More specifically, progress along the roadmap will require increasingly detailed decisions on exactly what is meant by "standardized and uniformly documented web services," and these specifications will be hollow unless they are widely understood to reflect official EarthCube positions. The implication is that the EarthCube decision makers must regularly and frequently inform and be informed by the technologists and scientist who assume responsibility for realizing this concept.

User Support - Even if specifications, as above, are officially sanctioned, they may be rarely or poorly utilized unless EarthCube includes systems that support deployment and use. For web services as envisaged in this roadmap (including attendant software libraries/frameworks/etc.), an adequate support system must exhibit long-term stability and likely would include a help desk, on-line forums, in-person forums (as part of an annual conference, e.g.), and perhaps a regularized set of user-training programs. Theoretically the building of an appropriate user-support system could be part of this roadmap, but it seems most economical to build a single, long-term-sustainable user-support mechanism that covers all aspects of EarthCube, so we treat the matter as a governance concern, at least for now. (While advocating a single support mechanism, we appreciate that it could be implemented in a distributed manner, with responsibilities assigned to numerous organizations.)

Resource Allocation - Support systems and decision making as above will carry no value unless there are adequate resources to make them effective. Simply stated, assignments of responsibility should be accompanied by assignments of resources (as well as authority), and such assignments are a fundamental aspect of governance. This is true for a number of EarthCube concepts, and it applies specifically to establishing a service-oriented architecture (built on web services) as envisaged here.

User Input and Feedback - Though the team or teams that implement this roadmap will have substantial amounts of end-user interaction, the effectiveness of the resulting capabilities will be further enhanced by user feedback from a broader base. Ideally, this should be amply coupled with user support, so we treat the matter as a governance concern, at least for now.

Community Contributions – While the SBIP-E team already brings substantial diversity and strength toward realizing the overall goals of the proposed activity, it is clearly recognized that the growing EarthCube community and the myriad activities in other EarthCube-sponsored projects bring significant expertise, capabilities, and developments that will need to leveraged and integrated into the design and development of EarthCube-related web services. As such, the governance process will need to ensure that such community contributions are not only welcomed but encouraged, facilitated, carefully reviewed, and integrated into the web services offerings.

Consistent with the above principles and with the aim of advancing the proposed SBIP-E activities, the management of the project will:

- **1** Establish a management structure, including working groups, and the governance process to carry out the proposed activities
- **2** Disseminate/Discuss Concepts and (Preliminary) Web-Services Specs
  - **a** Members of Our Own & Proximate Teams (Data Access, Cross-domain Interoperability, Workflows, etc.)
  - **b** All EarthCube Teams/Groups
- **2** Gain Input and Feedback on Prototype, Multi-Domain Realizations from
  - a. All EarthCube Teams/Groups

a. Representative Scientists in Multiple Disciplines

2. Disseminate/Discuss Specifications Revised to Reflect Feedback
    a. Members of Our Own & Proximate Teams (Data Access, Cross-domain Interoperability, Workflows, Etc.)
    b. Emerging EarthCube Groups

1. Develop Frameworks that Facilitate Realization (incl. “Long Tail” scientists)

a. Server and Client Libraries
b. Server Systems Targeting Big Data & Long Tail Deployments

1. Adopt Specifications as an EarthCube "Standard"
2. Establish Support Team/Process to Facilitate Large-Scale Deployment
3. Maintain Frameworks & Specifications at State of Art
4. Provide Relevant Training to Both the Users and Developers of Web Services

10. Risks

Google Doc catalog of services:
https://docs.google.com/document/d/16RxwV3kNoXs1DoDw_8CRZBfT9siT9JAepvNLZVTRakM/edit