A Service-Based Architectural Approach to Enhanced Data Access

Lyndon R. Oleson

U.S. Geological Survey
Center for Earth Resources Observation and Science
Sioux Falls, SD 57198 USA
EMail: oleson@usgs.gov

ABSTRACT

A common goal of most science data archives and information providers is to offer the widest possible access to their data and information products. In many science data management organizations, this responsibility can involve a number of heterogeneous legacy systems developed over time and in some cases tailored to a particular data collection, data type, or end-user community. Continued advances in the technology and functionality of the Internet are increasingly complicating an already challenging task for such data providers. Users and the tools and services they employ are becoming more sophisticated, resulting in heightened expectations for the functionality and performance of data discovery and access systems and services. Providers that develop and maintain their own data discovery and access clients or portals find themselves in competition with the rapidly evolving and expanding tools and services of the public Internet. For such providers, a more efficient alternative to attempting commensurate, incremental advances within current and legacy clients and portals is to make archived data more directly accessible to the larger community of Internet public domain discovery and access tools. One approach is to employ Web services and other interface standards to provide access to data servers by a variety of external clients and portals. This paper describes this service-based approach, explains some of its benefits, and identifies some of the limitations and potential implementation challenges of the approach.

INTRODUCTION

The U.S. Geological Survey Center for Earth Resources Observation and Science (USGS/EROS) is a science data management, systems development, and research center for the USGS Geography Discipline. USGS/EROS holds the world's largest civilian collection of space and aircraft imagery of the Earth's land surface, including over 30 million images representing over 2.5 petabytes of data acquired from satellites, and over 8 million aerial photographs. USGS/EROS is also a focal point for information about data acquired by other countries' Earth observing satellites. Resources at USGS/EROS include advanced data analysis capabilities, state-of-the-art digital data storage and archive facilities, production data processing systems, and online data and information access and ordering capabilities for USGS/EROS holdings and the land data holdings of other centers.

Since 1991, USGS/EROS has offered free access to metadata, or information about its holdings, by means of Internet connections evolving to Web-enabled access tools in the late 1990s. While these capabilities have been popular and successful through the years, it has become increasingly challenging to keep pace with the changing expectations of users. Enabled by advances in the technologies, functionality, and performance of personal computers and the Internet, users and the tools and services they employ are becoming more sophisticated, resulting in heightened expectations for the functionality and performance of data discovery and access systems and services. User communities are becoming more heterogeneous and fragmented with increasingly unique needs, thus data and information access system architecture drivers are becoming more granular and numerous.
Data providers like USGS/EROS that develop and maintain their own data discovery and access clients or portals find themselves in competition with the rapidly evolving and expanding tools and services of the public Internet. Therefore, data providers need to seek ways to achieve greater flexibility and adaptability to change, and to satisfy the expanding domain of users and tools, while minimizing the expense of revamping legacy systems. This paper explores the potential benefits of utilizing a Web services-based architectural approach to achieve these objectives.

**WHY EMPLOY WEB SERVICES?**

Many of the earth science data and information systems of the past 10-15 years were developed with tightly integrated Web-based clients and servers. Most servers interface with only one or a relative few clients probably developed by the same organization and tailored around a particular data type and set of user query and access functions supported by a tightly integrated server. These systems have generally been efficient and effective within themselves, but as Internet tools and functional models have evolved these systems have a difficult time keeping pace in such areas as scalability, performance, and reliability. In addition, users are seeking greater flexibility and efficiency in data and information discovery and access mechanisms. Many users now expect to be able to customize their own workflows. Users want to access many different data and information collections from a single, favorite client or application tool rather than having to move from data-provider client to data-provider client to query and access different collections. Tightly integrated client-server systems do not readily accommodate access from many different user clients. Interoperability across data systems and servers, whether within a data center or across organizations, is much more difficult to achieve. A service-oriented architecture, and in particular a Web services-based architecture, offers significant promise to addressing these problems.

A service-oriented architecture (SOA) is an information technology approach in which applications make use of loosely coupled services available over a network. The network can be an internal or intra-organizational enterprise service bus, or it can be the public Internet. When the World Wide Web serves as the network used, this is normally referred to as a Web services-based SOA. Loose coupling means that the client of a service doesn't have to know very much about the service to use it. This is particularly attractive when there is a need to functionally link heterogeneous, legacy systems with newly developed or evolving systems as is the case for many data providers like USGS/EROS. A Web-services based SOA, provides a way for the data providers to achieve broad systems interoperability in a highly flexible and effective manner.

The Web services-based SOA vision of interaction between clients and loosely-coupled services enables widespread interoperability across traditional boundaries (e.g. organizational, geographical, or system). The way a client communicates with the service doesn't depend on the implementation of the service. The client doesn't need to know what language the service is coded or on what platform the service runs. The client communicates with the service through a specified, well-defined interface, but then the service is free to satisfy the service request internally without needing to be aware of the technical nature or structure of the client.

For example, USGS/EROS users could employ a popular client to discover, obtain or analyze information through a suite of USGS/EROS advertised services without needing to know or understand the structures or mechanics of the various data and information systems that are satisfying the requests behind the scenes. On the other hand, by making data and information available through this same set of common or standard services, USGS/EROS data system servers are not required to know or understand the details of the myriad of user clients being employed to access them. This loose coupling via services allows users and USGS/EROS data and information systems and servers to functionally interact without the overwhelming complexity of creating tightly coupled, one-to-one links from each data system to each client. This approach also enables new clients or new data and information products and services to be added to the enterprise system without significant additional cost or delay associated with tightly integrating across all clients and servers. This architectural approach also allows data providers to offer
access to legacy systems through Web services interfaces or “wrappers” without needing to significantly change or alter the core capabilities of the legacy system, potentially allowing the data provider to offer much wider user community accessibility without having to offer countless locally developed and maintained clients. In this manner, the functional capabilities abstracted from the legacy systems that provide them enable the legacy systems to be more easily updated, migrated or decommissioned without significant, direct impact on the systems that interact with them.

OPPORTUNITIES AND CHALLENGES

Web services are not new to the USGS and USGS/EROS. Responding to trends in geographic information systems (GIS) and the desires of users to access and interact with large, seamless, geospatial datasets; the USGS released the Seamless Data Distribution System (SDDS) in January 2001. Historically, large geographic datasets were partitioned due to physical, hardware, or software limitations and sometimes political considerations. By removing arbitrary boundaries, seamless datasets provide researchers a more natural and powerful mechanism to analyze and study large areas of interest. The SDDS (http://seamless.usgs.gov) allows Internet users to determine areas of interest, establish output parameters, and directly download geospatial data. In addition to offering access to SDDS holdings via several “map viewer” clients, SDDS also offers Web services access to GIS-ready information through map services published on the USGS Web Mapping Portal (http://gisdata.usgs.net). Web based interoperability for seamless data holdings has been achieved through OGC’s Web Mapping Service since 2001. USGS/EROS has also extended the access functions of the SDDS to a Tiled Data Distribution System (TDDS) implementation and both SDDS and TDDS are in the process of being updated with Web services “wrappers” around the data extraction subsystems to offer expanded Web services functionality to users.

While SDDS has afforded USGS/EROS some significant experience with employing Web services in the delivery of both seamless and tiled map or georeferenced data sets, this capability and approach has not yet extended itself into the area of access to the vast satellite data collections of the Center. There are several reasons why steps to follow SDDS in employing Web services in the approach to access systems for satellite and aerial data collections have been slow, but the most significant reasons relate to the format and volumes of the data.

Historically, many of the satellite and aerial data collections have been maintained in a raw, unprocessed form. Partially due to the sheer volume of data involved, and partially due to the desire to offer users processing parameter options at the time of ordering, the vast majority of these satellite data are neither preprocessed nor available for direct download. Where SDDS focuses on directly downloadable map-registered, GIS-ready data, the systems that provide browse and order capabilities for these unprocessed satellite data are metadata-based “catalog” systems where users place orders for products to be generated before being sent via media or staged for subsequent download. In addition, the majority of the aerial data collections are analog film collections where products are also ordered on demand from a catalog of metadata.

While the on-demand catalog search and order paradigm has been the predominant business model for satellite and aerial data for many years, the evolution of the Internet and user access tools has allowed USGS/EROS and many other providers to move toward offering direct Internet access to preprocessed data products stored online. This direct data access paradigm lends itself much more readily to employing the kinds of Web services approaches employed by SDDS and other systems. Direct access can be provided, through a suite of common or standard Web services interfaces, directly to data servers via any number of compatible user clients and applications. The data provider might also maintain one or more clients of their own for various reasons, but these would not be the only clients available to the user community at large. Also, in addition to basic data downloading, extended services such as subsetting or reformatting or user driven processing could become readily available to users via the same Web services interfaces. Offering user driven processing services could be a key benefit allowing data providers to tap user momentum and ideas to hasten innovation.
Regardless of whether data or information are directly downloadable or orderable through a catalog interface, USGS/EROS sees benefit in employing the Web services-based SOA approach. Direct data and information access is the preferred functional model of both users and data providers, but in some situations direct data access may not be possible or feasible. Even in situations where catalog query and order interfaces might need to be maintained as the primary data access paradigm, benefits can be gained from employing a loosely-coupled interface between clients and servers.

The two primary Web-based catalog query and order capabilities presently available to USGS/EROS users are Earth Explorer and the Global Visualization Viewer. Earth Explorer (http://earthexplorer.usgs.gov) is a traditional catalog search and order system that allows users to query online catalogs of metadata and, in many cases, reduced resolution browse images of archived satellite and aerial data collections for selection and ordering. These data are in relative raw form with little correction processing having been applied and users are offered a variety of correction processing options when ordering. Once ordered, products are generated within hours or days depending on the data type and product being requested, and then are either staged for network download or sent directly to users on media. The Global Visualization Viewer (GloVis) (http://glovis.usgs.gov) is a complementary search and order tool that was designed to make maximum use of the low-resolution browse images for selected satellite data collections. Through a graphical map display, the user can select any area of interest and immediately view all available browse images within the USGS inventory for the specified location. From the browse image viewer page, the user may either navigate to view adjacent scene locations or select a new area of interest.

USGS/EROS is in the process of re-architecting and redesigning Earth Explorer and GloVis to take advantage of some of the benefits of a Web services-based SOA approach. Each of these systems is currently a tightly coupled client-server system where neither the client nor server functions of the systems operate independently from one another. This does not easily allow other clients to access the servers; users can access these catalogs and functions only via the dedicated Earth Explorer or GloVis Web sites. USGS/EROS plans to insert standard Web services interfaces between the client and server functions of Earth Explorer and GloVis. By taking these steps we see future users being able to query, browse, select and order these same products from any number of user-preferred clients as an alternative to the USGS/EROS-maintained Earth Explorer and GloVis clients as long as these user-preferred clients can interact with the Web services interfaces that are provided. These steps will also allow the client and the server functions to evolve and expand separately providing greater opportunity and flexibility to augment or enhance in the future. In the case of the client functions, it is hoped that this approach will relieve the stress of attempting to keep pace with the rapidly evolving public Internet geospatial and graphical discovery and access clients. In addition, we anticipate some cost savings from greater component reuse in the development and implementation of new data and information access systems, and these steps should help to facilitate the ongoing transition to directly downloadable data delivery options and enable expanded or enhanced services to be provided.

While there are many potential benefits to utilizing a Web services approach in systems such as Earth Explorer and GloVis, there are a number of practical challenges that need to be overcome to realize the full potential of the approach. For example, there have been some modest prototype efforts to provide access to the satellite data inventories behind Earth Explorer to external clients via Web services and, while the efforts have verified the technical approach, other practical problems are very evident. In 2006 a prototype Web service was developed in collaboration with the U.S. Department of Agriculture (USDA) to experiment with providing access to USGS/EROS satellite data catalogs by a USDA client. This service provided searches of the Landsat satellite data inventory by orbital path and row and (optionally) dates. The metadata was returned in an XML format that was specifically designed for the prototype. The experiment was technically successful and validated the Web services approach, but one of the practical performance considerations it revealed was the need to limit the number of catalog inventory records being returned.

Another of the practical challenges for a data provider offering public Web services is how best to advertise them. How will users learn what data information and Web services are available to them, and
how to access them? In the Web services environment these kinds of questions are answered via service registries. A service registry is a searchable online catalog that is utilized by a service provider to “publish” a service. Many service registry tools are available in the industry, but these will need to be sorted, selected and packaged in a way that is easily accessible and understandable to USGS/EROS users.

The overall performance of Web services is another significant area of challenge and concern. Depending on the characteristics of the data or information being served and the complexity of services associated with the data, many areas of potential performance challenges need to be considered and overcome. For example, the nature of the frequency and type of interaction between a client and a Web service can have a significant impact on overall performance. Network congestion or excessive memory and processor overload could result if clients need to issue multiple requests to a Web service to complete a single logical operation. Selecting an appropriate data transfer strategy, particularly when large amounts of data are involved, is critical to the performance of a Web service. Offering a service extension that performs compression and decompression or offloading data transfer to other services can potentially help performance. Data caching techniques are often employed to increase performance, but these can also be challenging to implement in certain combinations of data types and user demand patterns. Data caching can also involve a significant investment in direct-access disk technology when the size of the data granules being cached are relatively large as is the case with the satellite imagery offered by USGS/EROS.

As Web services become more popular, a critical problem that arises is managing the performance of services under extreme overload. Care needs to be taken to build highly scalable systems that have the ability to consistently serve the requests despite variations in the volume of requests. Such systems need to maintain a high level of throughput of Web service requests while minimizing the amount of latency between requests and responses. For large scale data and information systems, this performance concern involves much more than the overhead of handling and parsing service request. These concerns go inside the servers to the realm of data storage and retrieval mechanisms. Especially in cases where legacy systems are being accessed via Web services interfaces, the ability to scale and expand these systems may be limited. USGS/EROS is not only investigating and implementing ways to enhance the sheer performance of underlying storage and retrieval systems, but also exploring ways to control or “throttle” request during peak demand periods or to guard against malicious situations such as denial-of-service attacks. The SDDS has used “rights management” processes to assign priority to different user groups based on need or urgency of data access.

SUMMARY

USGS/EROS is seeking ways to achieve greater flexibility and adaptability to change in its online data and information access systems, and to satisfy the expanding domain of users and tools, while minimizing the expense of revamping legacy systems. A Web-services based SOA, provides a way for data providers to achieve broad systems interoperability in a highly flexible and effective manner. The loose coupling of clients and services allows users and USGS/EROS data and information systems and servers to functionally interact without the overwhelming complexity of creating tightly coupled, one-to-one links from each data system to each client. This architectural approach also allows data providers to offer access to legacy systems through Web services interfaces or “wrappers” without needing to significantly change or alter the core capabilities of the legacy system, potentially allowing the data provider to offer much wider user community accessibility without having to offer a myriad of locally developed and maintained clients. USGS/EROS is in the process of re-architecting and redesigning Earth Explorer and GloVis to take advantage of some of the benefits of a Web Services-based SOA approach.