EarthCube and Cyberinfrastructure for the Earth Sciences: Lessons and Perspective from OpenTopography

Christopher Crosby, San Diego Supercomputer Center
JRamon Arrowsmith, Arizona State University
Chaitan Baru, San Diego Supercomputer Center

INTRODUCTION
The National Science Foundation-funded OpenTopography Facility (http://www.opentopography.org) is an example of a project that is successfully developing production cyberinfrastructure that serves a diverse community of users. Through the leveraging of emerging cyberinfrastructure technologies such as portal-based data access, service oriented architectures, high-performance database systems, open source software libraries, optimized processing algorithms, and high-performance computing platforms, OpenTopography has dramatically improved internet-based access to massive high-resolution lidar topography datasets. OpenTopography's growing user community of several thousand scientists, educators, students, government agency staff, and private sector users illustrate that cyberinfrastructure-based geospatial data access systems can have a significant impact by democratizing access to these massive data sets.

OpenTopography's success is an illustration of the potential opportunities that exist through the application of cyberinfrastructure resources to geoscience data management and processing. With the vision and systems prototyped as an R&D project under the NSF GEON Grid project, OpenTopography also represents an excellent example of the process through which early cyberinfrastructure research efforts are evolved into production, stable and reliable, community resources.

DATA AND CHALLENGES
High-resolution observations of the Earth's surface, collected with airborne and terrestrial lidar (light detection and ranging), high resolution imagery (satellite and airborne), and multibeam bathymetry technology are revolutionizing the way we study the geomorphic, tectonic, biologic and anthropogenic processes acting along the Earth's surface. These data are fundamental tools for research on a variety of topics ranging from earthquake hazards to urban systems modeling. These data also have significant implications for earth science education and outreach because when visualized, they provide an accurate and detailed three dimensional digital representation of landforms, natural hazards and processes, and the built environment. Repeat measurements can be differenced to assess changes in the surface.

However, along with the potential of these data comes an increase in the volume and complexity of data that must be efficiently managed, archived, distributed, processed and integrated in order for them to be of use to the community. These massive datasets are often difficult to manage and poses significant distribution challenges when trying to allow access to the data for a large scientific user
community. Furthermore, the data can be technically challenging to work with and may require specific software and computing resources that are not readily available to many users.

TIERED DATA ACCESS
A key aspect of OpenTopography is a tiered approach to data access (see figure at left). In order to facilitate access to lidar data for users with varying ranges of expertise and computing resources, OpenTopography provides a variety of data products. At one end of the spectrum, expert users can access entire multi-terabyte point cloud datasets, often many billions of individual points, and run a particular processing routine with customized parameter settings. The system also permits users to produce derivative products such as common geomorphic metrics like slope, as well as visualizations of the products. This on-demand access to data and dynamically generated products is enabled through the co-location of raw data with the high-performance processing resources available at the San Diego Supercomputer Center. By leveraging remote data and processing co-location, users are able to explore and optimize data products for their scientific application.

In the middle of the spectrum, pre-computed raster data products (DEMs), at resolutions that are optimal to the datasets and organized into manageable tiles, are popular data products. These products are easily imported into desktop Geographic Information System (GIS) software packages, and require no LIDAR-specific knowledge to analyze.

At the other end of the spectrum, non-expert users can access pre-processed network-linked Google Earth images that provide streaming access to 100s of gigabytes of LIDAR-derived imagery for synoptic data browsing. These products enable non-expert users to quickly and seamlessly access large volumes of scientific data. The overwhelming usage of these data products demonstrates the impact of this simple yet novel approach for delivering easy to use lidar data visualizations to Earth scientists, students, and the general public.

SERVICE ORIENTED ARCHITECTURES
The technical goal of the OpenTopography Facility is to build a modular system that is flexible enough to support the growing needs of the Earth science LIDAR
community (figure below). In particular, we strive to host and provide access to datasets as soon as they become available, and also to expose greater application level functionalities to our end-users, e.g. generation of custom Digital Elevation Models (DEM$s$), and hydrological modeling algorithms for generation of derivative products. Furthermore, we envision enabling direct authenticated access to our back-end functionality through simple Web service APIs, so that users may access our data and compute resources via clients other than Web browsers.

Service Oriented Architectures (SOAs) have several features that are beneficial for building cyberinfrastructure solutions for science scientific data systems. Through encapsulation of back-end functionality, SOAs reduce the complexity of the processes that are part of the scientific workflow. Applications and data resources can be wrapped as Web services, exposing only a thin API to clients. The implementation details and the infrastructure back-end can be completely hidden from end-users and scaled up without any change to the service APIs. Web service APIs also enable interoperability across systems and architectures through the use of open standards, such as XML, SOAP and REST. Clients may be written in a variety of languages, and run on disparate platforms. Due to these benefits, we have implemented an SOA to accomplish our goal of making OpenTopography an open and scalable community resource.

THE ROLE OF OPEN SOURCE
Development of production community cyberinfrastructure systems must heavily leverage other ongoing software development efforts in the open source community. The OpenTopography system is built upon a (mostly) open source software stack, and many of the fundamental processing capabilities provided are dependent on open source libraries. Furthermore, open data and service standards (e.g., OGC) are also necessary to permit system interoperability and data integration. The SOA approach does permit us to run critical commercial applications within the workflow and for which we have yet to find or build an appropriate alternative.

CATALOG SERVICES
Metadata catalogs serve an important role in enabling discovery of data products and processing services. OpenTopography is in the process of implementing an OGC
CSW-compliant catalog that permits discovery of data hosted by OpenTopography and our federated data partners. Such a catalog service would return metadata about data products available from OpenTopography, but would also provide information about the processing capabilities offered by OpenTopography. For example, a bounding box query would inform a user that in addition to Google Earth image overlays, standard DEMs, and point cloud data products are available from OpenTopography, custom DEM products and derived products are also supported by OpenTopography. This catalog service vision can be extended to computational services that can be accessed independently of data. In the case of OpenTopography, various high performance gridding services are offered, as well as other independent services that can produce derived products and visualizations from DEMs.

DATA CURATION
Any cyberinfrastructure-based data system must provide robust data curation in order to ensure the longevity and reusability of data. OpenTopography is currently working with the University of California, San Diego Research Cyberinfrastructure Initiative to formalize our data curation and archive procedures. As part of this process, all incoming data products will be captured and archived to the San Diego Supercomputer Center cloud-based data storage system (https://cloud.sdsc.edu/), and issued a Digital Object Identifier (DOI) to ensure the data are persistent and citable. Data products dynamically generated by OpenTopography have their provenance captured through dynamically generated metadata that describes the processing algorithms and parameters applied. Ultimately we envision allowing users to issue a secondary DOI that describes their custom-generated data product, and that could be used to cite the specific data used in their scientific analysis.

USABILITY AND INTERFACE DESIGN
Although there are numerous informatics challenges associated with building the EarthCube, we believe that attention must always be paid to usability, interface design, and utilization of modern web-technologies. In a world where the average internet user can find the answer to nearly any question through a simple search box in the middle of a nearly white page, earth science data systems must strive for similar levels of simplicity and usability. OpenTopography has emphasized this theme throughout the development of our system and make a conscious effort to incorporate modern and familiar web technologies into our user interfaces (see figure below). Although scientific data present significantly more complexity in terms of parameters and query conditions, system builders need not immediately expose this complexity to users. Dynamic, Ajax-based interfaces make it possible to hide advanced query options and processing parameters from novice users, while still allowing advanced users to access the parameters they may wish to manipulate.
Further, role-based portal systems that adjust as a function of user type and permissions, and personalized workspaces that mimic the shopping cart environments that proliferate across e-commerce sites on the Internet also help to ease usability and make systems familiar and intuitive.

Our portal based system serves several thousand users per year, some of which are registered users accessing billions of lidar points, while others are guest users visiting to download Google Earth visualizations and lidar-related sources hosted on the site. In the past year, OpenTopography users have run ~4000 on-demand processing jobs accessing approximately 125 billion points. Beyond access to data, OpenTopography is also a hub to related training information, software tools (http://opentopography.org/tools), and documentation. We support our user community through short course training, responsive “customer support”, and an active social media presence (OT Blog, Twitter, and Facebook). OpenTopography has also begun to enable social interaction around data, permitting users to share their dynamically generated data products with colleagues and other members of the OpenTopography system.

CONCLUSIONS
The lessons above stem from the experience of migrating the OpenTopography lidar data access and processing system from an R&D project under the GEON Grid initiative to a development and production (D&P) system that serves a growing and vibrant community of earth science users. OpenTopography’s success is an illustration of the power of applying cyberinfrastructure resources in a practical manner to geoscience data management and processing, and the ability of such systems to democratize access to the next generation of earth science data products.