EarthCube DeepSearch:
Discovering and Integrating Data from the Geoscience Deep Web

Chaitan Baru, David Nadeau
San Diego Supercomputer Center, UC San Diego

I. Introduction

The NSF Workshop on *Envisioning a National Geoinformatics System for the United State*s, held in Denver, CO on March 14, 2007, described *"…a future in which someone can sit at a terminal and have easy access to vast stores of [geoscience] data of almost any kind, with the easy ability to visualize, analyze and model those data."*

This is more than just a vision. Indeed, such a capability is essential if future scientists are to conquer and make effective use of the vast amounts of data that we are producing, and will continue to produce, in the geosciences and elsewhere. There is a critical need for a "search engine" capability that makes discovery of data easy, and enables users to access the data in a convenient form for immediate use in downstream processing with other tools and software applications.

The current situation where scientists just "know" where to find the data that matches their needs, and "know" how to navigate the specific—and sometimes arcane—interfaces at each data provider site, is untenable. There will be too many sources of data and information in future. Future researchers and future graduate students cannot, and ought not to, spend much of their research hours learning all of these details of data access. With the amount and variety of data that will be available, this would leave relatively little time for actually working with the data to do novel research.

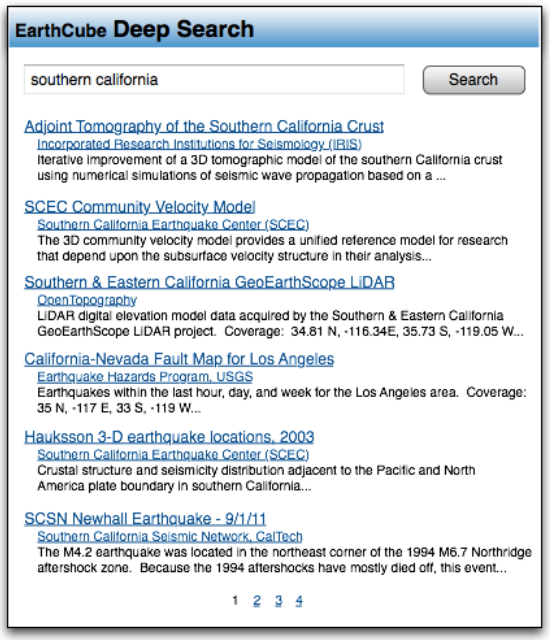
Just as the Google and Bing services enable easy access to information on the Web, a discovery service such as the proposed *EarthCube DeepSearch* is needed to enable easy access to the large and increasing amounts of geoscience data that are being generated and made available online. The *DeepSearch* service should be as easy and efficient to use as current Web search tools. It should provide access to the metadata associated with datasets as well as to the actual data itself, wherever possible.

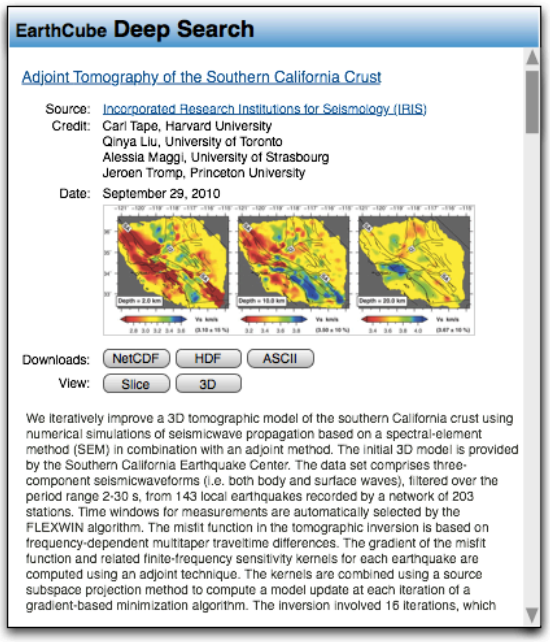
II. The EarthCube DeepSearch Service

EarthCube DeepSearch would be a "deep search" engine: it would search through metadata catalogs; navigate specific interfaces at well-known data archives; and process XML (sometimes HTML) formatted pages of metadata. It would be natively aware of the geoscience context, for example it would know about important websites for geoscience data; it would be aware of basic geospatial concepts, e.g. spatial extent, elevation, observed vs modeled vs simulated data, as well as basic geoscience data concepts, e.g. issues related to projection systems, dataset resolution, "raw" versus derived products, associated or related data products, and basic semantics of the domain.

The notional outputs of a DeepSearch are shown in the images below. For example, a search on the term "Southern California" might return a number of datasets related to that region such as, say, the "Adjoint Tomography of the Southern California Crust", the "SCEC Community Velocity Model", the "Hauksson 3-D earthquake locations 2003", etc. There might also be other datasets related to oceanographic and atmospheric data, though the user may also have the option of restricting the search context using terms such as, say, "solid earth" or "geology and geophysics". Clicking on one of the search results would show a more detailed page for that dataset, as indicated in the second image. In this case, for the "Adjoint Tomography of the Southern California Crust", the full, available metadata is provided, along with thumbnail images,

as well as buttons that allow access to the data in different formats, e.g. NetCDF, HDF, ASCII, 3D, etc. This capability may be provided natively by the original data source, or may require additional processing by DeepSearch to provide the specified output. The





set of buttons may vary depending upon the services offered by the data source and the type of data.

III. Related Efforts

There is a rich tradition of metadata catalogs in the geosciences. THREDDS catalogs abound among the community. Data archives like IRIS, UNAVCO, MGDS, CUAHSI HIS, CZO, and several others strive to provide web services and online interfaces to metadata catalogs. We are ourselves members of the US Geoinformatics Information Network (USGIN project, PI Lee Allison, Arizona Geological Survey) where we have implemented a prototype CSW catalog. USGIN is federating catalogs across all 50 state surveys and the USGS. In 2002, the GEON Project pioneered the notion of a semantically enabled metadata catalog (see http://www.geongrid.org) to capture information about "untethered" datasets, i.e. those that did not naturally fit into the mission of extant geoscience data archives. Projects such as OpenTopography.org, based at SDSC, will also provide metadata catalogs for their resources.

DeepSearch would be designed to navigate through all such resources—referred to as *search targets*—including well-known data archives and catalogs published by individual projects and/or PIs, to perform a search on behalf of the user.

Example: We use OpenTopography.org as an example to illustrate DeepSearch concepts. A user might issue a DeepSearch request by specifying a bounding box encompassing, say, a region in Southern California. DeepSearch would search through all of its search targets, with OpenTopography.org being one of them. For the given query, OpenTopography would return metadata about datasets that intersect the bounding box region. Along with this, it would also return information about what data products are available for that dataset. In the case of OpenTopography, data may be available as point clouds, DEMs, or KML files. For each data type, OpenTopography would return information about what processing could be performed with that data (which could then

be displayed as buttons, as shown in the above image). For example, DEMs would be associated with "download" buttons, while point clouds would be associated with processing buttons that would enable users to produce custom DEMs. Clicking on the different buttons would cause DeepSearch to invoke the corresponding function at the source and return the result to the user.

As a data search service, DeepSearch would also provide a *workspace service,* which provides users access to a workspace for temporary data. The results of a search, including downloaded data products, could be temporarily stored in such a workspace (and removed after a period of time). The concept of such a workspace was also used in the GEON project.

IV. Search Targets

As mentioned, DeepSearch would target "well-known" geoscience sites that have metadata and data. These obviously include the major NSF geoscience data archives, such as NCAR, IRIS, UNAVCO, MGDS, OOI, as well as other major facilities/projects, such as CUAHIS HIS, OpenTopography, CZO, and perhaps even catalogs from adjacent disciplines such as LTER and NEON. Any catalogs published via known, published interfaces—such as the CSW service from USGIN, various THREDDS catalogs accessible on the Web—would also be candidates. Catalogs from other agencies, such as USGS, NASA and NOAA would also be targets.

The search terms provided to DeepSearch could also be entered into a Google/Bing search, of course. While the results may be extensive (like a typical web search), they will not have the level of domain understanding of the geosciences that would be embedded in DeepSearch. Nonetheless, it would be possible to issue a simultaneous Web search and return those results in a separate window, if the user wants to do so.

V. Implementation

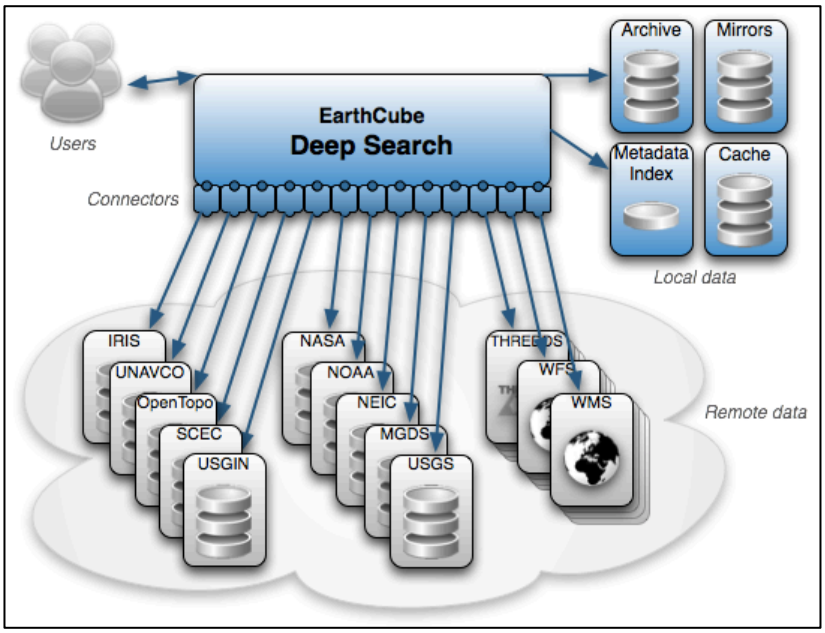
Over the years, we have had extensive experience implementing a number of metadata catalogs and portal-based search services. As mentioned, we pioneered the notion of a semantically enabled metadata catalog in the GEON project. The GEON catalog employed the DLESE metadata schema. Our group was also the lead developer of the EarthScope Data Portal (http://portal.earthscope.org), which employed the notion of a "Data Cart" and provided a user workspace for bundling together and storing search results. Recently, we implemented a prototype CSW catalog for the USGIN project, using an ISO schema standard. We are now in the process of implementing a metadata catalog service for the OpenTopography.org project.

We currently have an internally funded R&D project at SDSC entitled "Exploring PageRank for Data" (PI: Nadeau), where we are exploring the notions of search result ranking and "relevancy" for metadata and data search (as opposed to web search). Geosciences data is one of the use cases for that study.

V.1 Phased Implementation

The accompanying figure shows the software structure for DeepSearch.

The implementation would proceed in iterative phases. The first phase would be to identify the key search targets and develop search/crawl modules for each target. This will proceed concurrently with identifying the search terms structure and typical search phrases. DeepSearch will clearly have spatial, temporal, and keyword-based search dimensions. In the first phase of implementation, we will simply return hits from the identified sites and wherever possible return links to the actual datasets.



In the second phase, DeepSearch will return information about available services that can be invoked on the data. As mentioned before, these could be performed at the search target site, or by the DeepSearch service itself. The notion of a DataCart and workspace will also be implemented in this phase.

Our implementation will utilize both mirroring and caching for better search performance. Where we can get agreement from the search targets, we would actually mirror their catalogs at a central location, in order to speed up search. This would require a mechanism by which the mirrors are kept in synch with the original catalogs. Additionally, we would routinely employ caching strategies to also improve search performance. We also plan to implement DeepSearch in the cloud to utilize the distributed infrastructure of the cloud both for efficiency (localize processing to where the request was issued) and resiliency. The DeepSearch workspace as well as the mirrored and cached catalog information would be in the cloud.

VI. Partners

This effort will be open and highly collaborative with a broad set of partners, including all of the major geoscience data archives and projects that have data holdings. DeepSearch will work with existing interfaces, with minimal intrusion on the data providers sites/projects. We will reach out to organizations such as IRIS, UNAVCO, USGIN, CUAHSI HIS, NCAR, OOI, OpenTopography, etc. Any project that publishes metadata (via catalogs or otherwise) would be a candidate. The activity will be necessarily open and community-based. Solving the data discovery problem is in everyone's interest—for data providers, this will help drive more traffic to their sites; for data users, this will simplify their access to data.