EarthCube Charrette White Paper, October 2011

Earth Sciences: The Tower of Babel

And the Lord said, Behold, the people is one, and they have all one language; and this they begin to do: and now nothing will be restrained from them, which they have imagined to do. Go to, let us go down, and there confound their language, that they may not understand one another's speech. (Genesis 11:6-7)

Anna Kelbert
Oregon State University
Corvallis, OR, USA

Gary Pavlis
Indiana University
Bloomington, IN, USA

Abstract

The major challenge of Earth Sciences today is addressing the problems of general interest through intense collaboration that bridges disciplinary boundaries. Given the diversity and complexity of available data, such collaboration is particularly timely and would lead to a true leap in understanding. However, it is greatly complicated by the fact that the science has steadily diverged and evolved to the point of the Tower of Babel: scientists working on the same problem from different angles of view are having trouble communicating their ideas to each other, and sometimes are not even aware of each other's work. To bring this diverse knowledge base back together, we need to start at a common ground, which we suggest could be space-time location. We need an infrastructure that allows scientists to easily submit, and then pull up, information collected or computed in a particular area of the 3D Earth, at all scales, possibly as it evolved in time. Such products must become a standard output of Earth Science research along with peer-reviewed publications, and relevant metadata much accompany all such products. In this environment, fruitful interdisciplinary collaboration could easily flourish. Here, we suggest in broad terms a technical approach that would allow us to proceed with this idea.

Introduction

With the advent of EarthScope and other major recent Earth Science initiatives our community is awash in enormous quantities of data from these new large sensor networks. Within disciplinary boundaries, great progress is being made in understanding the Earth processes. The data are being used for geodynamic and inverse modeling to infer the composition and dynamics of the Earth's interior, as well as for more direct analysis. The final products of such research range from detailed three or four dimensional models of the Earth to qualitative conclusions, and anywhere in between.

This has produced a series of challenges:

1. The science is being strongly limited because people adapt existing algorithms and find that many are not feasible when extended to the newer, larger data sets. New methods are constantly being developed making use of the modern day rapid expansion in computational capabilities, but many of these have not yet matured enough.
2. Data processing is a completely fragmented mess with each group building complex, fragile data processing setups customized to fit each research problem. One group even gave their approach the playful name DUDE (Discovery Using Duct-tape Excessively) (http://www.iris.edu/hq/es_course/content/2011.html)
3. In addition, Earth models and related products derived from these data have a wide variety of formats, poor metadata content, and are not widely available for inter-disciplinary research. They are also becoming increasingly detailed with subtleties not adequately captured by traditional publication methods.

Although all three of these items are important issues that we hope EarthCube can address, this paper focuses mainly on item 3. Mutual understanding within the Earth Sciences is greatly complicated by the specifics of each of the sub-disciplines. The language and the methods have steadily diverged and evolved to a point that even neighboring Earth Science disciplines working to solve the same question are having difficulties communicating. Once results are obtained, they need to be preserved in their full complexity and interpreted in the context of all other geophysical knowledge. All components are already in place for a dramatic progress in the Earth Sciences except for making intense interdisciplinary collaboration part of public practice. Then, if geophysical knowledge continues to progress at its' current pace, it is theoretically conceivable that in some 20 years we shall understand and model the workings of the Earth in such fine detail that it will mark the end of Earth Sciences as we know it; some other discipline will take its place. However, a true leap in understanding will only come through a joint analysis of all the available data.

Although a similar challenge stands in the way of much of natural science nowadays, we argue that Earth Science in particular is fortunate in that it deals with a single object – specifically, the 3D Earth, possibly also evolving in time. We therefore suggest starting with space-time location as a common ground for communication between Earth scientists.

Developing a cyberinfrastructure framework for Earth Sciences

Research products in modern Earth Science include data, models, as well as qualitative conclusions and visualizations used to help others understand ideas. Most pertain to a specific area of the 3D Earth. Some apply everywhere and are therefore global. Some apply at a particular scale, which also vary enormously, from micro to kilometers to the full spherical Earth. If they were all collected into a database, a subset of these products could be pulled out by selecting a point, a line, 2D profile or a 3D area, and perhaps a scale and a time. Our long-term view of a cyberinfrastructure for Earth Sciences involves precisely such a database, equipped with appropriate visualization (aka Google Earth), selection, data and image processing, comparison and collaboration tools. A simple concept would be to think of it as a Geographical Information System that is four-dimensional. In other words, the Earth Science community needs to move from a focus on maps and cross-sections to treating all Earth Science data in a joint four-dimensional framework. Maps and cross-sections should be viewed in their proper context as sections for the native 4D objects.

As a first step towards this vision, we suggest focused development of two distinct products: a data framework and model framework, with the aim of merging them into a joint cyberinfrastructure in the years to come. An important step in this direction has been made by the IRIS DMC team in the last year or so: SPUD (http://www.iris.edu/spud/) is a first draft for the data products framework, and the EMC (http://www.iris.edu/dms/products/emc/) is an attempt to make it easier to submit, pull out and plot model information. Both of these approaches mostly address a first-order component of the problem commonly called data discovery. That is, they are pieces of software a data center produces to allow scientists to find and obtain the data they need more easily. That is, however, only one part of a much bigger problem. We also note that GEON (http://www.geongrid.org) addressed many of the issues raised here, but never quite got to a comprehensive solution. Another example includes GeoMapApp (http://www.geomapapp.org/), but that only addresses the problem of 2D geophysical data visualization. We suggest collaboration with IRIS DMC and other interested teams to achieve the following first-order goals.

1. A robust suite of three-dimensional visualization tools that can be used easily with Earth Science data. This has two elements: web-based data exploration tools and software to run on a local machine. Both are necessary and currently completely divergent. Every web site has a different look a feel and there are dozens of competing commercial and open-source packages directed at 3D visualization. The problem is that none of them mesh well and scientists spend a great deal of time with format conversions and writing ugly unix shell scripts to assemble all the data for a 3D scene. We need a clean path from data archives like IRIS to desktop visualization systems that are up to the task of handling modern, detailed models.
2. Tools need to be developed from the ground up using an object-oriented approach with simple interfaces to common scripting languages like python and perl. A reasonable design study is needed, but most of what we need can probably be derived from an existing framework such as the

Visualization Tool Kit (http://www.vtk.org/). In a sense what we need is standard geometric objects that know where on Earth they are and, when necessary, what time they existed in that state.

3. Operations to simplify the process of comparing 2, 3, and 4D models need to be developed. An extension of common image processing operations would also be developed. For example, a user could pull up two models showing different physical quantities in the Earth, and use our envisioned tool to run a range of "comparison" operations, checking for correlations or anti-correlations both in the models, and in their gradients.
4. Communication tools such as a basic social network or chat should be an integrated part of the infrastructure, facilitating remote communication with colleagues and allowing to share the images / screen to aid a scientific discussion.
5. Many (if not all) of the communities in the Earth Sciences are using a variety of historic or self-derived file formats for their products. Many of these do not have adequate metadata component and are peculiar to a niche field. This makes it difficult to maintain the information in the years to come, and even to share it within a certain disciplinary group. Common metadata content needs to be specified that encapsulates key concepts without being encumbered with the details of a particular specialized fields. For example, 3D Earth models could be viewed as a three-dimensional field object. Many things one field may view as essential are metadata to another. For example, a tomography model might have a link to a resolution test while a model produced by a finite difference simulation would need the boundary conditions applied to that simulation. Development of extensible and self-describing file formats by each of the geosciences communities would constitute a major part of the proposed effort.

Conclusions

The Earth Science community needs a comprehensive four-dimensional, object-oriented system that will allow scientists to easily interact with and compare various digital data products. A simple vision of the idea is a Geographical Information System that goes from 2D to 4D, and which interfaces cleanly with Earth Science data centers. This would require a joint cyberinfrastructure that crosses sub-disciplinary boundaries across the Earth Sciences, and the development of well-defined, self-describing formats for information storage.

The proposed cyberinfrastructure would provide a framework for maintenance of the valuable geophysical information in the future years, as well as for sharing and interdisciplinary collaboration. It would be a critical and timely development, allowing for a revolutionary progress in the Earth Sciences.