EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

(Jennifer Arrigo, CUAHSI and Ying Fan Reinfelder, Rutgers: January 29-30, 2013)

Earth Cube Workshop Title: Hydrology/Dark Data – Envisioning a Digital Crust

Introduction: This workshop brought together geoscientists to develop a community vision and a path forward to achieve a "Digital Crust" – a three-dimensional digital representation of the composition and structure of the continental crust of North America that would advance our ability to quantitatively describe, model and understand fluid flow in the subsurface, from the critical zone to the deeper crust. While the primary background of participants was hydrology, and those interested in modeling water flow in the critical zone at local to global scales, some participants came from other geosciences and research areas such as geophysics, geothermal energy, stratigraphy, geochemistry, and geodynamics/deep crustal fluid flow. There were 43 on site participants, and 18 virtual participants via WebEx, mostly (74%) from US based institutions. The stakeholder survey indicated that many of the participants were experienced researchers (74% having over 11 years of experience in the field). Main outcomes from the workshop are listed below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years (list 3 to 6).

   - A high priority is understanding the evolution and functioning of the earth's critical zone, defined as the thin near-surface layer of the crust that sustains all terrestrial life. Fluid circulation and thus enabled energy, carbon, nutrient and other geochemical fluxes play a critical role in shaping the evolution of terrestrial biosphere and societies. The structure of shallow groundwater flowpaths, and its exchange with surface waters and the vegetation root-zone determine the seasonal water availability to vegetation and aquatic ecosystems, as well as carbon and nitrogen transformation and transport. There is no information on the material properties below the soil survey depth (~1m), preventing interpretation of field observations and modeling efforts across cm to watershed to regional scales.

   - Another high priority for our science is to advance a synthetic understanding of forcing of groundwater flow over many scales. Currently groundwater assessments are done at discrete scales, and information is not typically transferred between scales (upscaling or downscaling). The digital crust effort could provide a means to evaluate forcing of groundwater over a very wide range of scales (local, regional, continental), and to understand linkages between scales (e.g., effects of changing precipitation patterns and sea level on regional-to-continental groundwater levels, cumulative effects of water withdrawals, effects of regional-scale modifications of land use and surface drainage networks), as well as provide the basis for better incorporating groundwater in earth system models in ways that allow us to evaluate two-way feedbacks between groundwater and climate system on much larger and longer timescales then currently possible.

   - Another fundamental science question that bridges several geosciences disciplines and has extreme relevance for society is understanding the role of fluids in seismicity and tectonics.

How can we quantify the distribution and magnitude of fluxes from the brittle to the ductile regime; can we better under the interaction between the hydrosphere and lithosphere?

- Share different interpretations of available data into geologic structures. Data standards and tools are not currently adequate to allow domain scientists to share interpretations or to quantitatively compare and contrast different interpretations of the various kinds of geologic, geophysical, and mineralogical data used to infer geologic structures.

- Organize the variability, connectivity, averaging, and covariance of disparate physical and chemical properties of the crust within the context of geologic structures. One of the central challenges identified by having workshop participants discuss their knowledge and challenges within their own disciplines is the fact that the earth appears to have many structures depending on the particular properties used to define the structure, but many applications require synthesizing information on multiple properties (e.g., weathering → temperature, mineralogy, water flow and chemistry, etc.; nitrogen dynamics → temperature, water flow, oxygen, carbon, microbial community structure, active microbial biomass and/or metabolism). At the current time, we don't have a good sense of how these various representations of earth's structure compare, or at what scales different properties average, or how important properties co-vary (or don't co-vary). We also need to directly face the fact that all estimates of structure have a high degree of uncertainty.

- Advancing our understanding of paleo-reconstructions of depositional environments. A specific example discussed at the workshop was the Gulf Coast. Building a complex model of the 3D geology of depositional environments over several periods of time (from Mesozoic to present day). Mapping these over time gives a much better understanding of the complex stratigraphy of these depositional environments, allows targeted sampling of geologic features to derive source evolution for tectonic investigations, and can aid societal needs such as energy exploration.

- Another high priority science challenge identified by some participants is to further our understanding what the geologic, geomorphic, and environmental factors that determine the formation of the unique environments – e.g. karst systems. Karst systems exemplify the type of transformative and societal important research the digital crust would enable. Karst covers about 20% of the earth's surface, and are incredible fragile environments - subsidence (natural hazard), water quality, and urban/other planning that needs to understand the impacts of karst geology on water supply, construction, and other issues. Researchers currently studying karst environments often have to create their own datasets from many disparate and regional sources, and these studies often create a wealth of data that are not easily shared, so we do not have a comprehensive picture of global karst research and information.

2. **Current challenges to high-impact, interdisciplinary science:** Several themes emerged as consistent challenges faced within/across the involved discipline(s) (list 3 to 6).

   - The major theme that emerged is that in all these applications we are dealing with a poorly observable system. The constant challenge is how to represent different interpretations and capture the knowledge of the scientists that created them, and to quantify and deal with uncertainty. In a discussion focused on primarily hydrologically-relevant properties, it was discussed that there could (would) be multiple geovolume interpretations. Sharing knowledge across disciplines would require substantial metadata or context. A central hypothesis of this effort is that a reference geologic framework will be useful to organize attribute data;

simplifications of this geologic framework can be done for different hydrologic or geochemical applications.

- Another challenge in dealing with an poorly observable system is that most data are inferential: for example, geophysical logs can be interpreted to indicate geologic formations but inversion algorithms are subjective. The type of data used to even create a simple 2D or 3D geologic map includes several data types (well logs, seismic reflection data, samples, published and paper references) that must be integrated.
- A central challenge in creating the digital crust is the issue of dark data. Much of the data investigators are using for their research are not digitized. The baseline subsurface information for much of the continent is sparse. Many participants stressed the amount of time and effort it took to track down and assemble suitable data. In addition, many researchers lamented the fact that once their project was completed, there was no way to share the information they had collected. In fact, as an anecdote, two participants discussed their research on areas within the Chesapeake Bay. These researchers were literally working in areas within 100 miles of each other, but were working with different agencies and sources, and each spent much time tracking down, assembling and digitizing data. And each lamented that after doing all of their work, there was no easy place to deposit or share the new data resources they had created.
- Another challenge identified was the disparate repositories for the type of data needed to assemble a digital crust. The US Geological Survey has the most comprehensive data resources and has a goal of building a 3D geology for the nation. There are also state geological surveys with data (both digital and non-digitized) that could inform the digital crust. In fact, many participants at the workshop relayed developing relationships and working directly with state geological survey personnel, and employing graduate students to discover, digitize and format data from these offices. If we add in data being created, assembled and/or used by the geophysics and other geosciences communities, there is a clear need for governance and coordination.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

1. **Desired tools, databases, etc. needed for pursuing key science questions with brief elaboration:**

   - We envision a database composed of a collection of fundamental geologic units – including, but not limited to hydrostratigraphy and soil horizons. The system should accommodate these units as 3D GeoVolumes; this system should allow the size and shape of these geovolumes to evolve over time. We would envision a "reference" set of geovolumes, governed and maintained jointly by the academic community and the USGS that represent the consensus best available continental 3D geology.

   - The system should be able to represent multiple interpretations or sets of geovolumes – a way to think about this might be the way geodatabases can have multiple layers that contain different representations or interpretations of surface properties. A user may come into the system through the continental 3D geology geovolumes but could then access other researchers' interpretations of geovolumes over a specific area, find local or regional studies that have more detailed or high resolution information, etc.

- The system will need to contain and present substantial metadata in a way that allows both expert and non-expert uses to evaluate the interpretations and geovolumes for their quality, appropriateness, and fitness for use in different applications or models. *This was seen as a central, unresolved challenge by the workshop participants – communicating uncertainty, transferring the inherent knowledge, context and understanding of the scientist who makes the original interpretation, etc – are all key*

- The system must have an easy way for researcher to share and deposit their own data. The system must have ways for researchers to not only share their own data but to feedback to current data in the system – e.g. a researcher might contribute high resolution data set on a particular reason – this data should then be incorporated to our larger understanding of the system, and could/should result in a change in the "reference" set of geovolumes size and geometry over this area. This would require oversight/governance system to be set up.

- The system should have a way to represent and share proprietary or protected information (e.g. metadata only). Many researchers relayed experiences of working with data that is proprietary. Participants felt it was important that the digital crust convey the existence of this information as well as contact information for people to request access.

- Behind each geovolume requires a provenance, i.e., comprehensive archive of all supporting data and sources. Users could access this archive and work with the data directly to create their own geovolumes, extract data of interest, etc. The data system would have to accommodate variable resolution in x,y,z for the data underlying the geovolumes. The data system may have to accommodate gaps or "no-data" geovolumes.

- The domain of this data system would be from the land surface down to where data is available and material definable (a "goal" could be the brittle-ductile transition). The data system should easily integrate with other data systems as much as possible (e.g. surface data, DEMs, vegetation, etc. so that there are not mis-matches or discontinuities) so that researchers could easily assemble data needed to investigate critical zone or earth system processes.

- The data system should support a suite of data retrieval and analysis tools, allowing users to explore and access the data flexibly. Specific examples the workshop participants cited:
  - Flexible selection of spatial domain, grid resolution, generation of x-sections and geo-volumes
  - Enhanced visualization and ability to "video fly-thru" such as done by Google Earth; integration with other data sets. An example was given of viewing Google Earth or a DEM, and then having the ability to "peel back" the surface and see the subsurface underneath.
  - Algorithms to calculate grid cell properties (different means, std dev, functional forms, etc).
  - Ability to generate 3D grids of specific material properties (physical/mechanical, chemical, biological)
  - Ability to incorporate uncertainties or probabilities in 3D location. A specific example was researchers who wanted to create 3D GIS features of specific geologic features (e.g. sand bodies, areas of a specific threshold of an important property, e.g. high or low permeability) but wanted to be able to represent the uncertainty in the location of these features (since they are interpreted) – the system could create a 3D grid of probabilities of whether a feature was present, and 3D features that could represent specific probability thresholds as concentric shells.

- Although logical data models exist for representing 3-D geologic formations, the current tool set for working with 3-D geovolumes is inadequate to domain scientists. Standards for serving and exchanging such geovolumes are nascent at best.

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward how it can use EarthCube to achieve those goals:**

    - Development of a generalized 3-D geologic map of North America is possible and would provide a useful starting point for developing the cyberinfrastructure necessary to maintain, evolve, and utilize the Digital Crust.
    -

    • Achieving that goal requires maintaining and growing the community of researchers working in this area and establishing some initial communications/community resources to share data and experiences. We can do this initially with the workshop participants and the Google Docs site to share research, resources, presentations, and thoughts. EarthCube could support these data communities through the Ning site, working groups, further workshops, to maintain momentum and coordination.

    • Expertise from current Earthcube groups (in particular, the groups looking at interoperability and semantics) could be utilized in working groups that could look more in depth at developing metadata models and standards to address the concerns of workshop participants given the uncertain, interpreted data products.

    • EarthCube will be needed to develop, implement and maintain the community governance needed to realize the digital crust as envisioned. Such a system would need to be jointly maintained by NSF (representing the academic community) and federal agencies (such as USGS).