Technological Grand Challenges

Compiled from EarthCube End-User Workshop Executive Summaries

Last updated 24 November 2014

Contents

Introduction

In order to reach out to potential end-users of the National Science Foundation's (NSF) EarthCube initiative, NSF funded a series of two dozen EarthCube domain end-user workshops throughout mid-2012 to late 2013, targeting a broad spectrum of Earth, atmosphere, ocean, and related scientists, including senior and early career scientists. The purpose of these workshops was to allow geoscience communities to articulate and document their cyberinfrastructure needs and what they would like to do in the future, in terms of accessing data and information within and outside their disciplines.

A specific goal of these workshops was to gather requirements on EarthCube science-drivers, data utilities, user-interfaces, modeling software, tools, and other needs so that EarthCube can be designed to help geoscientists more easily do the science they want and need to do. More specifically, science that helps address the NSF's GEO Vision, 2009: fostering a sustainable future through a better understanding of our complex and changing planet.

This document is a collection of all the technological grand challenges identified during these end-user workshops. Workshop participants were asked to identify key technological challenges they face in data acquisition, comparison, integration, curation, etc. in pursuing key science questions, as well as the desired tools, databases, etc. required.

Articulating Cyberinfrastructure Needs of the Ocean Ecosystem Dynamics Community

1. Tools:
    - Tools to enhance data discovery and searches.
    - Visualization tools for interactive data analysis.
    - Tools to track and control data versions.
    - Tools to foster data quality assurance and quality control.
    - A community-level interface and facility to share tools and programming code.
    - Technology to assimilate metadata produced by smart sensors.
    - A tool to translate from different format types to a standard format.
    - Automatic incorporation of new data into databases/repositories
    - Automatic retrieval of data for use in applications (e.g., forecasting)

2. Repositories and Databases:
    - New data repositories (or expansion of existing facilities) are needed for emerging data streams that currently are not supported by a repository (e.g., metabolomics, citizen science data)
    - A system for handling massive amounts of data including images and video.
    - Production of a wider range of more sophisticated data products and derived calculations.
    - Ensuring repositories function as, or work with archive facilities for long-term preservation of data.

3. Global Infrastructure:
    - Guidance on where to submit data including restrictions and guidance for repository use.
    - A centralized forum providing information about models, scripts, software, and documentation.
    - Undergraduate and graduate-level curricula for training the next-generation of scientists to be able to find, submit, and work with data.

- Educate programmers to understand science.
- Standard interdisciplinary metadata format.
- Cross-domain ontologies of measurable phenomena and instrument types.
- Further development of crowd-sourcing funding and technology

Bringing Geochronology into the EarthCube Framework

1. Geochronological data are currently difficult to access, of variable quality, and challenging to compare between labs/methods and with other information;

2. Limited standardization of data acquisition, archiving and delivery protocols across the geochronologic community;

3. The need to archive legacy data and develop mechanisms for managing the current data explosion, so that new and existing data can be leveraged;

4. Domain-specific data architectures are vastly incomplete or absent altogether and require the development and maintenance of software designed for the reduction and archiving of geochronologic data, designed to remain flexible for unanticipated data additions;

5. There is a lack of transformative technology for integrating earth system knowledge -- data at present are locked in domain-specific architectures;

6. It is difficult to recognize gaps and data deserts in existing datasets, and disconnects between disparate datasets;

7. To create geochronology data that is amalgamated into databases directly comparable requires financial support of EARTHTIME-like initiatives for various geochronometers to establish community-wide protocols, and evaluate and improve inter-laboratory comparison;

8. It is challenging to develop ***continuums*** across human and geologic timescales;

9. Educational content for EarthCube users that includes support for preparing the next generation of geochronologists to benefit from EarthCube's "big data world";

10. A need for visualization tools that make EarthCube accessible to the non-specialist audience (e.g., non-specialist scientists, K-12 teachers, policy makers).

Community Based Cyberinfrastructure for Petrology, Geochemistry, and Volcanology

1. **Discoverability:** Improved infrastructure (search engines, catalogs) is needed that facilitates discovery of data, samples, models, and management tools.

2. **Interoperability:** Software packages utilized during data acquisition should be transparent to data analysis and visualization softwares. Tools should support analytical thinking and numericisms.

3. **Compliance:** Data and metadata should be captured at the point of acquisition in a way that they can seamlessly be managed throughout their life cycle, including upload to repositories in order to satisfy data management requirements.

4. **Format standards:** Standards need to be established within the existent data repositories.

5. **Sample tracking:** Systems should be in place to promote spatial contextualization of analysis through sample registration, imagery, and links between samples (hand samples, thin sections, splits, etc.) and analytical data.

6. **Archiving:** Absent repositories, databases, and heterogeneous media should be identified and recovered.

7. **Metadata:** Ancillary contextual information such as science objectives, data provenance, and uncertainty estimates at each step in workflow needs to be included with the data.

Community Modeling

1. **Algorithm development:** In parallel to the advances in computational hardware power, advances in algorithms, software, and compilers enable better, more effective use of advanced computing. Optimal algorithms become more critical as we solve larger problems on larger computers. Continued advances require support for developing portable mathematical and numerical methodologies across fields of geoscience. New methods require research by applied mathematicians, computational scientists, and statisticians (among others) that is motivated by geoscience problems.

   In addition to research advances, implementation of new algorithms requires skilled, experienced software engineers to develop and support community codes and assist geoscience researchers with code development. However, it can be difficult to recruit and support software engineers in the domain sciences; it is essential that attractive career paths and sustained support be available to talented software developers. The challenges and barriers to new algorithms include sustained support for both ends of this spectrum (research, and code development and hardening.)

2. **Visualization:** Scientific visualization is an essential element of the scientific work for modeling. Models can generate very large, complex, and high dimensional data; scientific visualization is a fundamental tool for analysis of these data, extraction of features, data assimilation, verification and validation of numerical methods, and extracting insight. Scientific visualization is used as a preprocessing aid to assemble inputs and discretization for models. Finally, scientific visualization is used to communicate results and discoveries to the research community and beyond, to policy makers, educators, and the general public. The technical challenges and issues include availability of adequate methods for visualizing complex and diverse data types, integration of visualization at all appropriate steps in the workflow, visualization of very large datasets, and adaptation of new technologies.

3. **Models:** Infrastructure is needed to support model reproducibility, reusability and transparency. Community models require sustained development and support and community tools for working with them, such as workflows and software for managing the enormous amount of scientific and computational choices that go into models. Community standards for testing, computing and portability of model codes would greatly enhance the impact of these models. These standards

would aid in the creation of more flexible and easier to use community models, and would enable more effective science in a research environment that has a rapid pace of scientific and technological development, limited resources for developing and sustaining meaningful collaborations, and an existing and enormous diversity in model structures, programming languages, computational platforms and data requirements. Such models should seamlessly access data resources and parameters.

4. **Advanced computing:** Modeling typically requires access to advanced computing resources, including (but not limited to) large-scale high performance computers such as are available from the Yellowstone- NCAR-Wyoming facility, NSF's XSEDE facility, and leadership class DOE computers. Advanced computing may also include mesoscale parallel computing, from small clusters operated by individual PIs to mid-sized clusters; these can be difficult for PIs to obtain and operate. Modeling science requires effective access to and assistance using such computing facilities, in order to make best use of the investments in computing hardware. New technologies (such as GPUs) are emerging, requiring redevelopment of models to take advantage of increases in performance.

5. **Model and Data uncertainty:** As multi-disciplinary efforts emerge to model multi-scale and long-term processes, researchers are challenged to identify systematic and rigorous ways to rapidly assimilate new data and to characterize the statistical structure of observational data. It is important to pay attention to systematic, random, and model error as well as possible sources of unknown errors. For even well-understood systems, predictive modeling with quantified uncertainty and model-based experimental design places new demands on characterization of uncertainty in both observational data and models.

   For less well-understood systems, different approaches must be explored. These different sources of error and uncertainty are not currently well-communicated, and to the extent that such communication takes place, it is usually only within a community or scientific domain, and not beyond. Communication of uncertainty is especially important for those who must try to craft policy from science. Since uncertainty quantification is an active area of research containing many open theoretical, methodological, and algorithmic questions, one challenge is ensuring that methodology and cyberinfrastructure be made extensible in order to support future innovations.

Cyberinfrastructure for Paleogeoscience

1. Intuitive 4D access to all existing knowledge products and underlying data/metadata and methods used to generate those knowledge products

2. Intuitive 4D mapping and visualization capability across data resources/model outputs

3. Better access to and discovery of tools and methods to manipulate and analyze data of different common types (e.g., time series, stratigraphic position)
4. Improved agreement upon standards and semantics for basic, widely-used data/methods, particularly for age/time representation

5. System for determining and dynamically updating age models (and uncertainties) within and between existing resources and model results

6. Improved user workflow and explicit reward system for data generators (e.g., acquisition and submission to databases/repositories

7. Coupled earth-life system models that have good two-way, "live" integration with distributed data resources

8. Increased awareness/utilization of existing resources within and outside of the paleo community and funding to sustain and improve these resources

9. Improved metrics to evaluate success and contributions of existing efforts on a community and individual level; metrics to evaluate successes of new efforts

10. System to identify gaps in existing data sets and prioritize/incentivize verification of contradictory information, as well as filling gaps with new records

11. New educational capability that is built upon data and results drawn "live" from existing resources

12. Support for long-term archiving and retrieval of digital data/tools and physical samples

13. Need 4D visualization(s) for researchers (easy data comparison, discovery of gaps, etc,), scientists outside the paleo community, educators, policy-makers, and the public

14. Legacy and dark data incorporation – noisy signal processing (sort out bad data, dropped data, sparse matrix data (missing images, geochem, geomorph; this type of tool is also fairly standard but it needs to be incorporated).

Deep Sea Dynamics & Processes

1. **Training and Awareness:** Many members of this community recognize that they are supported by existing data management efforts, and clearly stated that they do not want EarthCube to "reinvent the wheel". That said, there is insufficient awareness of and access to existing tools and infrastructure - including data contribution and data discovery tools, open source software, visualization tools, and data analysis systems.
    - New tools need to be developed to improve both data management and data analysis.
        - Particularly, there is a lack of tools that lessen the "burden" of data management and could be embedded in our scientific and daily workflows. New tools that allow for easier and earlier integration of data management activities within the workflow are essential for future data acquisition.
    - There is a large personnel gap in the community between data producers and data managers that could be resolved by facilitating training within the community to lower barriers to available tools and resources.
    - There is significant and well-founded concern that the community lacks sufficient resources for data preparation and that those efforts are not sufficiently recognized and rewarded. While infrastructure for citing data and has been established within several data systems (e.g. Data DOIs), nearly all professional citations continue to be focused exclusively on publications. Much of the hard work of data acquisition, data management, metadata production, and data integration is largely unrewarded, lowering the incentive for proper data acquisition and curation and increasing the gap between data scientists

and discipline scientists.

2. **Data comparison and integration:** Datasets are often not fully comparable *because*:
    - Metadata are incomplete and inconsistent;
    - Navigational precision is problematic across deep submergence vehicles. It is essential that exact locations (x, y, z, t) are precisely identified for each sample;
    - Foci differ from project to project. Improving mechanisms for pre-expeditionary communication and developing tools to enhance collaboration (either at particular sites or for particular types of sampling projects) would maximize project utility and drastically increase funding efficiency;
    - Data formats and entry vary from project to project. This can be resolved with either format standardization or, preferably, algorithms that identify and correct for variation in format;
    - There is a lack (or a lack of awareness) of standardized methodologies to document sampling conditions, e.g., consistent definition of time stamps and locations for samples and measurements.
    - Data quality is poorly documented making data use outside the original research group and integration of disparate data sets inconsistent

3. **Desired Tools**

    Collaborative Tools

    - Tools are needed to facilitate real-time collaboration before, during, and after cruises. These include live ship-to-shore feeds that enhance shore-based participation in sample collection and real-time data analysis. Thus, expedition goals could be dynamic and responsive to real-time data analysis. Furthermore the use of Ancillary Project Letters (APLs) or RAPID-type funding models would allow for interested parties (this could focus on early career scientists) to join expeditions (in person or remotely) to collect co-registered or associated data/samples, thus increasing expedition efficiency. This is important for field-going scientists and modelers alike. This would also lower the barrier for early career scientists to undertake sea-going research by allowing for smaller projects to be funded and completed prior to pursuing larger expedition funding.
    - Mechanisms to better communicate caveats and built in assumptions necessary for interpreting data and models. Models, especially, need to continue to be linked to scientific expertise.
    - "Alert" system that will notify the user of a new data submission of interest. This could be developed to include not only data acquisition updates, but also self- populating personal databases and subsequent data analysis. For example, if one were interested in a particular metabolic functional gene in hydrothermal environments, a search/analysis/model algorithm could regularly self-update and new function gene trees would be the product for the end user. This goes beyond data discovery, but also automates data analysis, allowing scientists to focus on data interpretation
    - Experimental design, communication/ cooperation with various deep sea and related scientific communities

    Data Documentation Tools

    - There is a lack of tools (desktop, tablets, in the field (ships, ROVs etc)) that facilitate data

documentation and capturing metadata that can be used broadly by our community. This is a critical gap that needs to be filled if we are to effectively and efficiently feed content into EarthCube.

- We also need improved and expanded metadata and standardized metadata templates that easily identify units and commonalities (e.g. when, where (projection, coordinates), how (methods of collection, analysis), experimental design). Furthermore, we need to develop easy tools and simple guidelines for easily capturing metadata contemporaneously at the time of data acquisition.
- Data quality is inconsistent - EarthCube should include consistent and rigorous mechanisms for objectively documenting and evaluating data quality.

Visualization and modeling tools

- Many existing tools require extensive training for effective use or are incomplete. This not only inhibits usage across our community, but also limits our ability to analyze legacy data or integrate and analyze disparate data sets.
- EarthCube should include a clear and well-organized user interface with a well-documented set of modeling and visualization tools (with training documents) that can be improved or extended in modular form.
- We need more data integration tools, including tools that easily allow you to merge cross-disciplinary data (different data types) and tools that allow users to look at multiple data sets on a global scale. One idea was: "EarthClip" (J. Smith) - Integrated digital (desktop) guidance to help you discover data, contribute data, comment on data quality, etc. (e.g. "You may also be interested in…").
    - Easily accessible interface for using open source tools, without requiring installation on individual computers – cloud based, web page, all-encompassing application.
    - Tools needed for interactive figures (3-D) for both processed and raw data.
- Current data sets are enormous and the volume and quantity of data is only going to increase (e.g. HD video is becoming the norm, acoustic datasets, and someday (soon) biologists will be sequencing entire genomes for every organism in a sample). Moving these data sets will be (and is now) an enormous challenge and current solutions are rather antiquated (e.g. we currently ship large hard drives around the globe in order to share data and collaborate on interdisciplinary projects). We need to transition to cloud-based platforms that allow analyses in the cloud with systems that are connected with ultra-high bandwidth networks.

4. **Data Curation and Access Issues/Challenges**

- Relational databases that discern both user interest and intent from search parameters are now common in ecommerce, and could be applied to scientific data searches. For example, when you search for a spatula on Amazon, it shows you a bunch of other spatulas that other users also looked at. Is there a way to have EarthCube know or learn from users about connections between datasets in order to improve data discovery?
- Access to legacy data is important but is often difficult - EarthCube should include legacy data and/or clear links to legacy data, including ways to objectively evaluate the quality of legacy data. Incorporating legacy data into EarthCube is essential for maximizing its impact in the deep sea science community, however this community will only buy into this platform if there

is guaranteed longevity.

- Lost data sets as well as data sets that don't get pushed into the public domain are not uncommon. We as a community need to continue to be vigilant about data compliance. Can EarthCube make it easier to find and upload data into various databases? Can it be a two way street? Tools that lower the barrier between publications and data upload and curation to data repositories are essential in order to minimize lost data sets and ensure compliance with funding agency requirements for data management.
- Reducing barriers to access include cross-directorate, cross-agency data linkages ("Data without borders"). This includes NIH-NSF cross communication, potentially combining geological data with 'omics data. Public and private as well as national and international agencies (e.g., ONR, Schmidt, Moore, NOAA, IODP, etc.) support deep sea data acquisition, making data multi-jurisdiction but there is no jurisdiction to the seafloor.
- Broad-based, interdisciplinary seafloor models and data sets need to be integrated with surface and coastal models, ideally by incorporating all of these in the EarthCube platform.

Developing a Community Vision of Cyberinfrastructure Needs for Coral Reef Systems Science

There were four distinct classes of community needs: (1) Databases and Portals, (2) Data Processing, Modeling and Visualization, (3) Education and Training, and (4) Internationalization.

1. **Databases and Portals:**
   - Desirable features include standardization (formats, collections, and representation), richness of metadata, and built on existing efforts with tools to create and query data across repositories that includes standard reference keywords, DOIs, and appropriate credit for data provider. Data integration should support imaging, sequence, environmental sensor data, and local observational data and delivery of streaming real time data from sensors networks. Quality of data and metadata is critical since web applications and interfaces may involve with time.
   - Needs to include links/connections to existing resources through a curated data portal that could cluster databases/data by types. The system should allow grass roots contributions through user data entry as well as information filtering and discipline or research theme by sub-setting. Digital tools should also provide a dynamic, collaborative workspace for a variety of sub-disciplines (bioinformatics, ecological studies, genomics, mathematical biology, programming, etc.). Identifying a funding strategy for sustainable data curation is critical.
   - There are many "dark data" or challenging data types (such as imagery or sequence data) that could be better managed and harvested from unpublished studies, secondary reports, desk drawers, personal collections, original data from earlier publications not archived that require a various types of standards and integration methods. Improved methods to deal with these data types may arise from industry-academic technology sharing translated to the coral reef research community.

2. **Data Processing, Modeling, Visualization:**
   - A system of data processing pipelines for bioinformatics/omics data for computationally intensive analysis tasks is critically needed especially those that take advantage of HPC resources (XSEDE). The KEPLER scientific workflow system is a potential tool to utilize.

These approaches would include traditional statistical modeling approaches, machine learning, and geospatial analysis and multimedia analysis for image/video/audio analysis and information extraction.

- We could visualize disparate data (space/time) with "easy to use" software tools (vs. immersive environments) that support online visualization simulations with user-directed parameters. Google Earth is a reasonable model for the portal interface. The visualizations can be used to communicate directly with public through interactive and applied community engagement. Maintain data and software version control will tools such as GIT or Mercurial.
- Improved software tools (such as API's: application programming interfaces) for linking ecological and -omics software packages are needed with open standards that facilitate coupling through modular-based software or middleware to connect processes; better interfaces for communication among software models; Free Open Source Software (FOSS); Glue code and provenance: automated metadata extraction/provenance from digital objects and (e.g., OpenDAP, HDF (file format): geodata, temporal metadata; Integration/ alignment: scale; measurement equivalence, reward coders and nurture new type of coral reef scientist/hacker
- Develop a coral reef simulation system that merges model components (forecast, climate change), is applicable to many locations; 3D; modular, with case studies; the WRF is an example (Weather Research and Forecasting)

3. Education and Training (human resources/workforce)

- The support of a cyberinfrastructure tactical team to support training, scientific programming, and database administration would help facilitate many of the data analysis, education, and training needs. The coral reef domain aware CI team could rotate between thematic-based resources and help with challenging projects to achieve scientific end products that non-CI researchers would have difficulty creating alone. This role may also be fulfilled by a "Campus Champions" such as a grad student or similar to connect geoscientists and computer scientists.
- The coral reef community would support web-based workshops on portals for data; data integration; data management and mechanisms for local group participation (e.g., web-based). Various topics for education programs were proposed including programming, data management; online repositories/versioning; imaging technologies/analysis; tools for temporal/spatial scaling; communication with managers; promotion of cross-disciplinary training and research.
- University programs in coral reef sciences may institute computer programming requirements in curricula or develop and offer a Certificate Program in Coral Reef Informatics to encourage cross-disciplinary training (between geoscience and computer science)

4. Internationalization

- Need to improve access and use of international data; issues include need to share data to improve our understanding of reef globally to promote international cooperation on global scale reef studies. We would like to connect people, institutions, government management to democratize research and get many contributors to interpret science and results.

Key Science Drivers/ Questions in Coral Reef Science: Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years:

1. What processes are relevant to understanding the biological responses of coral reefs to biotic and abiotic drivers across temporal and spatial scales?

2. What are the mechanisms of coral reef adaptation and acclimatization to climate change?

3. How does symbiosis influence the biology and ecology of coral reef organisms?

4. How does the abundance and diversity of coral reef organisms influence community resilience at local, regional, and global scales?

5. How will invasive species, disease, and parasites disrupt coral reef ecosystem structure and function?

Early Career Strategic Visioning Workshop

Desired tools, databases, etc. needed for pursuing key science questions with brief elaboration:
One-stop shopping for improved access to data, ease of sharing data, with standardization, and ease in citing data – a "closed circle" from data production, use, review, and publication:

1. Better funding of data storage options, with aligned ontologies, able to "keep up" with "big data," and capture of legacy and archival data

2. Minimizing time collating data and maximizing time doing science

3. "Hindcast" and predictive modeling capabilities

4. International access and access to the general public

5. Achieving the multi-disciplinary potential, with integration across fields, databases, and agencies, and an overall cultural shift

EarthScope

1. Deployment of simple web services across several domains with a sophisticated brokering system(s).

2. Workflow with standard interfaces to the underlying components. Initial prototype and then produce script

3. Report, propagate, visualization of uncertainties. Tools for model validation and assessment (misfits) of available models. "misfit comparison between these models is …"

4. Community driven evaluation of data

5. Enhanced access to underlying models and data in publication. Possibly require submission of datasets to repository as part of publication. (include workflow/scripts associated with data) "Carrots and sticks" to encourage sharing easy uploading. publishers/checkbook requires uploading

6. An enhanced, but simplified, open source GoCAD for the 3D and 4D spatial integration of geological volumes, points, lines, and surfaces. An Earth model construction environments that allows information to be correlated and plotted will extracting and calculating additional quantities. Visualization tools for 3D and 4D datasets. Comparing different 3D models quantitatively. Common framework to share and compare models. This should allow one to overlay different data/models geographically integrate petrological models and data seismic model library

7. A web-accessible plate tectonic reconstruction system (paleo-GIS) that allows the earth to be typified by a hierarchy in the scales of deformation, from global rigid plate motions, to regional deformation with local faulting along with paleogeographic reconstructions is required.

8. Data analysis software packages that are well connected to the data center(s). An important component would be standard data mining and pattern matching routines

9. Extend usefulness of the Computational Infrastructure for Geodynamics (CIG) for EarthScope-science. This could involve the development of an environment for inverse problems. • Much greater collaboration between cyberinfrastructure development in EarthScope and earthquake early warning activities

10. Deployment of robust hosting services with a distributed architecture. The system would need a rich privacy and permission control over content to facilitate sharing or restricting by users and user groups (sophisticated content management system). Provide repository for data and policy designed to encourage/coerce sharing that data.

11. Funding for domain experts to collaborate with computer scientists. Allocated funding which provides support for researchers who need programmers and a pool of "certified" EarthCube programmers. In general, I think development of new tools should be driven by the users. The most effective approach is to have programmers clean up tools initially developed by individual researchers to make them user-friendly. There are too many examples of tools initiated by programmers that end up being of limited use to the community because they were ill conceived from the start. “My main limitation is time to learn to use tools that are already out there.” 2) A virtual institute for community software in seismic and MT (and related fields)

12. Training and documentation for software and data centers. Involve users in the documentation process. Wiki or living document, user forums, and social networks are ideal mediums for communicating with the user base.

13. Standards APIs for querying 3-D seismic velocity models and flexible data structures for their representation that facilitate large models (10^8-10^10 grid points) and fast querying. Web portals for simple querying of 3-D seismic velocity models are needed to provide earthquake engineers with the parameters they need for simple ground-motion prediction models. Techniques for representation of epistemic uncertainty and small scale features in 3-D seismic velocity models and 3-D fault models.

14. We need to be able to access as much geoscience data as possible through the "cloud" and in the field. This requires vertical integration of datasets, where information is sorted/queried by location.

15. Easier connection between tomographic models and wave propagation codes would be helpful. If IRIS EMC is the standard, then we need to adapt our codes to readily read in these models. 2.

"Push-button" assessment of tomographic models, based on running a suite of independent earthquake simulations, then calculating various misfit measures.

16. Powerful client side applications to access diverse datasets in user-specific ways to conduct analysis and visualization (MATLAB would be a good substrate for this) Use-case oriented term and units translations between diverse datasets as is applicable to specific studies, represented as ontologies. Uniform, open, REST-oriented web services with domain specific terms and data, but tools to promote translation to other areas of study, as opposed to being simply homogenized to the lowest common denominator.

17. 3D seismic forward modeling with setup tailored by user - geographically indexed database with results from all geosciences and links to journal papers in ISI - cross disciplinary data access (at stages between raw data and published results)

18. Convenient and flexible interface for students to browse and manipulate seismic data. 3D seismic wave simulations at continental scale and periods <20 s.

19. On-demand processing environment to produce higher level products from raw SAR data (interferograms, InSAR time series). 2. Archive of higher level InSAR products

20. A user-governed model based on a simple API with data discovery and visualization capabilities in the 4D space, which would allow users to submit their own data, models and workflows. All version controlled and semantically enabled. 2. Scripts with a clearly defined syntax that could immediately make any program part of the system: one could select of subset of information, "click" and execute a workflow. In this setting, codes could be run by a user on selected data directly on the relevant remote servers, through the API. 3. Submission of research products to EarthCube could be part of the NSF data policy, while sharing controls could be set by users on their personal content. Contributions should be optionally peer-reviewed, citable, and author-tagged through an underlying social network.

21. Metadata descriptors that allow facile query of databases database exploration tools fast networks for transferring large quantities of data

Education

1. The undergraduate geoscience education community makes uses of a very wide variety of geoscience data types. To see the range and depth of data in current use in geoscience education, please browse the following collections: *Using Data in the Classroom: Data Sources and Tools: (*http://serc.carleton.edu/usingdata/resources.html) and *Earth Exploration Tool-Book Chapters:* (http://serc.carleton.edu/eet/chapters.html).

2. Cyberinfrasture desired by the education workshop (see further detail in full report):
    - Data germane to society's pressing problems
    - Field data
    - Ability to ingest and display student collected data
    - Tiered approach to data quality (allowing quality student data to be added, while keeping out inadequate data)
    - Near real-time data, and also historical archival data
    - Local informants' eye witness accounts

- Novice-friendly interface options, scaling gradually up to the full professional interface
- Support for data exploration and "making 'failure' cheap"
- Collaboration tools
- Supports for understanding uncertainty in data and model output
- Comprehensive and comprehensible metadata
- Simple and well-documented versions of geoscience models
- Support for student building of models

3. Pedagogical & social infrastructure desired by the education workshop includes:
    - Support for citizen science
    - Mentoring for both teachers and students
    - Assessment techniques for student mastery of data and modeling practices.
    - Tutorials and training sessions
    - Venues in which to share and build a community of practice
    - Support for diverse populations, including learners with disabilities, urban youth with limited access to nature, and adult professionals crossing fields.
    - Support for entrepreneurial enterprises

Engaging the Atmospheric Cloud/Aerosol/Composition Community

1. **Data Access Challenges**
    - Different types of users need different types of support (some, for example in developing
    - countries, just want to import data into Excel)
    - Cross-community access is the biggest challenge (within an domain community, it is generally
    - understood how to work with data formats and tools)
    - Need to understand the data characteristics (quality, provenance)
    - Enable scientists to find relevant, reliable data regardless where the data are archived and obtain the data in the form specified by the scientists
        - Search by location, date, topics, etc.
        - On-line services for providing automated data customization
        - Globally available data; data in other country's agencies
        - Able to integrate data from different platforms and repositories
        - System interoperability (inter-agency and international sharing)
        - Better metadata and standards for data understanding and usage
    - Can EarthCube provide a better search engine tailored for communities?

2. **Non-domain Understanding**
    - Challenge is in having users understand the uncertainty and errors associated with data.
    - Documentation is critical. Needs to be understood by others outside of the immediate discipline that created the data.
    - Data processes change over time

- Need education/training about data (perhaps a service EarthCube could provide?)
- Can EarthCube fund training for cross-disciplinary data management and informatics?

3. **Supporting small datasets in EarthCube**
   - Many groups, (e.g., research labs) have small, individually maintained datasets and do not have a large infrastructure to support them in the publication and management of them
   - EarthCube should support these small datasets
   - One example, might be an EarthCube sponsored cloud data management and publication service to simplify the process for smaller groups

4. **Supporting Data Management**
   - Challenge is in funding data management - for example, many research groups don't have the funding or time to do metadata creation.
   - Interest in an EarthCube supported cloud service or tools
   - Employ people within EarthCube who have library science and similar skills to help organize and provide access to data

5. **Data Quality and Standards**
   - Need a set of EarthCube recommended standards and best practices to facilitate the interoperability and sharing.
   - Bad data needs to be flagged
   - Need a rating system to help determine and convey what is the quality and type of published data.
   - User need to understand when they're using data at their own risk or when it is peer reviewed
   - Need mechanisms to catch uncertainties and errors in data before and after they are published
   - Unique dataset IDs are created to link datasets to publications and datasets to each other,
   - Suggestion to only provide a dataset an ID if it is peer reviewed and determined to be acceptable.
   - Should provide supporting documentation that describes how dataset was derived (algorithms, software used, etc.). And need to track data processes as they change over time.

6. **Supporting modeling and integrated analysis**
   - On-line data integration and analysis services
   - Tools and services to manage, archive, and disseminate model outputs for facilitating modeling comparison
   - Sensor-model coupling for facilitating model verification and validation with observation data
   - Sensor web and models as services

7. **Merging Existing and New Infrastructures**
   - Need to transition existing systems to EarthCube, not requiring an overhaul of existing systems
   - Need translators, converters, adaptors
   - Strive for common standards and practices where most effective

Engaging the Critical Zone Community to Bridge Long Tail Science with Big Data

General cyber-challenges include:

1. CZ data is diverse and much of it is "dark". There is no one-stop shop for even knowing what is available, let alone accessing it.

2. One constraint that limits community access to Essential Terrestrial Variables (ETVs) for watershed modeling is that the data sit on many servers, with multiple (and heterogeneous) formats, very large files, and complex security, making it difficult for scientists or students to use the data. A second challenge is that even if the above problems were fixed, the scale of the data and the tools necessary for data mining, fusion, and visualization are not yet readily available or usable by scientists. The problem of accessing and sharing real-time data collected by CZO scientists is a theme in this challenge.

3. Modeling, computation, and numerical prediction is carried out in an ad hoc manner with limited cross-domain collaboration (water-bio-rock) and without the benefit of close interaction with cyber scientists and numerical analysts. An outcome is that such results both challenging to obtain and are not easily reproducible.

Specific scientific challenges that require cyber-solutions:

4. Understanding diverse scientific worksflows by CZ scientists and applying appropriate tools to promote shared discovery requires a fundamentally new approach to how the scientific process will evolve from experimental data, to interpretation and models, to the creation of knowledge and wisdom.

5. Uncertainty and variability are fundamental to all CZ use cases. Across a range of activities -- from field experimentation where sensors are impacted by environmental noise, to issues of communication in wireless sensor networks, to real-time data assimilation in nonlinear spatially distributed models, to data and model analytics, visualization and computational steering -- uncertainty and variability must be addressed. Although these areas are effectively dealt with in individual CZ disciplines, there is not at present a general framework to efficiently deal with this specific challenge.

6. Closed technologies such as WSN's (wireless sensor networks) have evolved as proprietary products that are not yet useful for the Critical Zone problems where low-power, integrated, heterogeneous, co-located systems of research-grade sensors are necessary to resolve multi-state, multi-process discovery within fully coupled bio-geochemical hydrological systems. In particular, research-grade, low-power bio- and chemosensors are particularly missing in the integrated measurements at CZO's.

Envisioning a Digital Crust for Simulating Continental Scale Subsurface Fluid Flow in Earth System Models

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

1. 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.

2. 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.
3. 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*

4. 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.

5. 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.

6. 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.

7. 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.

8. 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 geovolumes
    - 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 nascent at best.

Experimental Stratigraphy

1. Many of the identified needs and challenges bridge the gap between scientifically technical issues and cultural issues of our scientific community. While many technical issues could be addressed at the individual investigator level, a community-scale effort would likely result in greater efficiency, and it is paramount for creating lasting cultural solutions. Most all needs identified below fit into this framework.

2. Cultural: Difficulties with incentives for data sharing
    - Intensive in time and financial resources to produce an effective data sharing platform
    - No rewards for this kind of investment. Possible reward types could include
    - Institutional support, i.e. recognition during tenure process
    - Incentives from NSF for public data availability or reuse
    - Recognition for data generation distinguished from interpretation in literature
    - For long-term monitoring a funded investigator should still get to work exclusively with new data on interpretations before sharing
    - Lack of long term solution disincentives investment in resource
    - Scientists spend time on science and not on management

3. Technical: Need for expertise in data issues within our community
    - No expert resources to call on for guidance and assistance in management
    - Need training for data management for students, etc. from the beginning of the project
    - Many institutions do not offer IT support to investigators

4. Considerations for international cooperation
    - Language is a problem. For example, programming comments all in Japanese may be easier for the individual Japanese investigator but may create challenges in sharing code.
    - International agreement on sharing of data is an issue.
    - Coordination of physical and financial resources for hosting data is an issue.

5. Opportunities to put our discussed ideas into action and test
    - Testbed data sharing site could be served in a distributive manner with single front-end combining data, metadata, models, etc.
    - Trial solution for linking documentation with data
    - Individual investigators can work independently to test differing solutions allowing faster discovery of models that do not work well. Compels move to more open source solutions.
    - Allows for growth of merit-based solutions in data/metadata structuring, i.e. not prescribed by committee but determined by acceptance and use
    - Opportunity to test if this expedites secondary use of data to answer broader scientific questions
    - More funded community discussions to further advance a plan for data storage and dissemination. i.e.: We have not gotten it all done in two days but current ideas can be verified or discarded.

Integrating Real-Time Data into the EarthCube Framework

1. Improved community infrastructure: access to improved communication infrastructure, on-demand computing and protocols for data exchange

2. Metadata generation for real-time data streams and tracking of provenance

3. Real-time signal processing, calibration, and quality control: existence of standardized software libraries

4. Real-time computing: software that provides the ability to process, produce, and transmit derived products in real time.

5. Tools for integrating and assimilating real-time observations: from differing geospatial and temporal resolutions

6. Playback tools for re-creation and analysis of phenomena and the observed environment of past experiments

7. Frameworks and secure mechanisms for remote operation of instruments

8. Real-time visualization of observations made at different temporal and spatial scales

9. Data discovery and access including data subsetting of large bandwidth streams

10. Rendering of observations with widely different time scales for real-time displays

11. Decision support tools and integration with tools for emergency management

12. Engine/middleware/platform that will combine all these capabilities for the community

13. Developing networks for dissemination - including social media, apps, and user driven interfaces/portals including citizen science, crowd sourcing and open data access

14. Mechanisms to discover software and hardware for real-time acquisition and processing and to provide guidelines in implementing real-time capabilities (e.g., SUB/PUB real-time streams, buffering data for remote access, etc.), education.

Integrating the Inland-Waters Geochemistry, Biogeochemistry and Fluvial Sedimentology Communities

Current challenges to high-impact, interdisciplinary science:

1. Addressing the various types of data: At a point (across space, these will become); Spatial (local regional global); Temporal (minutes, days, weeks/months, annual, decadal, geologic scale); Field Samples; Modeled Samples

2. Datasets that bridge measures of quantity and composition of complex constituent mixtures

3. Disconnect between continuous measurements and concentration data.

4. Challenge of observing and characterizing hot spots and hot moments (fluxes and processes that are highly concentrated spatially and temporally)
5. Using, sharing, and coupling broader models (geochemistry, hydrology, etc)

6. Challenges finding data:
    - Lots of free data, very different formatting, provenance, other data characteristics.
    - Zero order challenge is knowing what data is out there, knowing what resources provide access, etc.
    - Data discovery is lacking
    - Challenge in finding linked data - spatially and temporally
    - Downscaling and upscaling data
    - What are the core capabilities that end-user domain scientists need in terms of data management/cyber-infrastructure?
    - Can we filter content based on assumed associated data?
    - Community standards for data management would make our science practices more efficient. Low transaction cost is necessary for adoption.
    - Need benchmark data (such as for training/validating models). Model intercomparison portals, testing our model quality.

7. Challenges using data:
    - Data quality varies, across soil types, DEM, sensors precision, accuracy
    - Enormous datasets that are difficult to use
    - Units are different, and metadata doesn't always provide clarity
    - Curation. Heterogeneous data quality. Lack of information about quality. Reviews of data sets? Summary of data quality and characteristics?

- Clearinghouse function of Earthcube? Meta DataBase
- Vocabulary variation. Semantic search.

8. What are the pressure points for the community? AGU and Nature Geo saying "we won't accept this unless you link your publication to data". But...where does it go? Let's go to a couple of the heavy hitters and ask them to be the bad guys. Others will follow.

9. Leverage other, related disciplines for their solutions to similar problems. Which atmospheric/ oceanic lessons would be translated to our domain?

10. Need to change culture about code sharing. What are the incentives? Why do we have different rules for code vs data sharing?

11. How do we decide the scale for making decisions? E.g., OGC deciding on standards vs grassroots community?

Desired tools, databases, etc. needed for pursuing key science questions.

12. National data set from water treatment plants and sewage treatment plants (they generally are apprehensive and don't share but have great data)

13. International data sets (some countries do not share)

14. Smoothed county level data

15. Improved and standardized statistical approaches for small systems

16. High resolution data throughout the hydrological cycle, not just field season campaigns and communication about current projects, research activities

17. Tools for coupling models are important and useful.

18. Rating datasets and models. Need to understand relative value of data and models faster.

19. Need standards for data exchange and formats.

20. Understanding and curating what has been done and what could be available.

21. Digitizing the wealth of information that exists behind us (historical data).

22. A large, comprehensive catalog? Consistent formatting. Need crosswalk for vocab.

23. Central retrieval system; centralized searching, not necessarily hosting data physically

24. We need mobile science apps to make field work more efficient. Improved models of concentration discharge relationships (*how do we share models?)*

25. Maps of built infrastructure information and data (tile drainage, pipeline, sewage treatment outlets, past land use)

26. Ground water chemistry database

27. Continuous categories of soil maps and soil chemistry

28. Fertilizer use data

29. Watershed activity for research

30. Historical maps of land use, lead deposition etc.

31. Species distributions of fish, invertebrates, amphibians, native, invasive species

32. Hyporheic flow paths

33. Soil moisture maps

Meetings of Young Researchers in Earth Science (MYRES) V: The Sedimentary Record of Landscape Dynamics

1. There was a quickly realized consensus that a Google-Earth-like data clearinghouse would be tremendously helpful. This would be a place where existing community datasets could be searched both by topic/index and geographically (as well as temporally, for historical and stratigraphic data). (For example a search for "suspended sediment" and "discharge" might return datasets from the USGS, the Army Corps of Engineers, individual PIs, and local/state and international agencies.) Ideally, once desirable datasets are identified, a researcher could then download them in a similar file format/structure. Physical and numerical modeling results could also be included and geographically cross-referenced by the lab of origin and a specific location (if a model were related to a field case, for example), and model codes could be shared (in a similar manner to what is currently done through CSDMS).

2. Participants agreed that Google Earth (or similar intuitive geospatial interface) itself would be a desirable backdrop for this type of community resource and there was no need to reinvent the wheel in terms of user interface, for example. Access to linked datasets via a large, searchable data clearinghouse (where file downloading and metadata storage were reasonably uniform) would also help improve access and usability of disparate datasets. This might mean, for example, that a researcher would only need to learn one data upload/download system, which would empower users to access geological, geophysical, biological, and climatology data, for example, via the same interface, rather than having to learn a new protocol to access data from each discipline.

3. A centralized data clearinghouse would also provide a place for PIs collecting new data to upload their results, thereby blending both existing databases and accommodating the needs of researchers who currently rely on ad-hoc arrangements to store and share their data. Although NSF's new data-sharing requirements are separate from EarthCube, participants expressed concern that if new data acquisition/sharing isn't incorporated into the EarthCube model, some of the problems and challenges listed in section 2 will persist.

4. There was also strong interest in the suggestion that resources be allocated to digitizing and updating legacy data that is not currently available in digital form. Participants were very enthusiastic about this idea and felt that it would be a high-yielding investment.

5. The “universal Earth-science database” concept generated the most excitement among participants. Participants were less concerned with visualization and modeling resources, in part because it seems that existing software and collaborative websites (e.g., CSDMS) are suitable for accomplishing important research goals (or at least are not viewed as significant barriers to progress), although improvements in visualization software and access to expensive software licenses (particularly for evaluating LIDAR and seismic data) would be helpful. Ultimately, challenges locating, accessing, formatting, and compiling data are currently frustrating and stifling to participants in this workshop.

Ocean ‘omics science and technology cyberinfrastructure: current challenges and future requirements

1. Omic database development is required for curation, maintenance and data standardization that will allow for easy data submission, extraction and query. As well, tools for rapid and simple data query and metadata association are necessary. This includes federation with non---sequence---based datasets (e.g. metabolomics and lipidomics) into existing/emerging oceanographic 'omics database/analysis/visualization platforms. Environmental 'omic databases need to be: (1) federated (i.e., all datasets are interoperably queriable and transparently accessible), (2) curated (validated and updated, as for example NCBI nr datasets), (3) sustained (i.e. a five---year commitment of support is not sufficient), and importantly, (4) intuitively accessible to a broad range of scientists, and the public.

2. The ocean 'omics community would benefit from “Google---like” or “Kayak---like” search and suggestion functions/engines, that could query across complex and heterogeneous, federated environmental, oceanographic and 'omic databases.

3. Tools and mechanisms are required for access to high performance computing and statistical analyses of large scale 'omic datasets that could accommodate both naïve users as well as experienced “power users”. One possibility is a user facility that functions similarly to UNOLS oceanographic facilities, that would provide access to software developers, bioinformaticians, and analytical tools, as well as the hardware required (storage facilities, servers, clouds, etc.) required for 'omic analyses. Researchers could request access to this facility in association with successful grant applications, as with UNOLS. Extending the capabilities of BCO---DMO or similar services also seems another tractable model.

4. The community would benefit from access to a web clearing house/portal with links to standard “ocean 'omics” best practices, algorithms, software and workflows, as well as analytical and statistical methods under development, with entry points for both naïve and power users, would be a useful resource for the community.

Rock Deformation and Mineral Physics Research

1. Central data system for DEFORM and COMPRES science. This should include storage, visualization, and search protocols to provide community access to our data and solutions that will reduce activation energy to including data in these databases.

2. Community technical forums, including websites, focused on CI developments for both DEFORM and COMPRES. We note that COMPRES has a Technology Office at Argonne National Laboratory and a technology-oriented website maintained by COMPTECH, which could be a starting point. We need both tools to compare data from different labs, including functional fits, statistical analysis and model evaluation and a social network associated with our data system to provide a forum to interact virtually and lower barriers to interdisciplinary interaction between researchers. These tools should include a way to capture information about how users interact with the databases, and automated methods to improve the data system based on this information.

3. Central archive of experimental samples with integrated workflows, database templates, and community-wide DOI system for samples

4. Automated system for storage and evaluation of microstructural images, including rock fabric, texture evaluation and pore networks, as well as comparison of laboratory and field microstructures and shear zone texture.

5. Extending data mining capabilities/tools and interlinking existing repositories, (e.g. crystal structure and spectroscopic databases) with newly developed databases.

6. Reliable, sufficiently automated, easily accessible and well-documented software for efficient (preferably real time) processing of large volumes of experimental data and results from theoretical and numerical studies.
7. Improve accessibility of high-performance computing (HPC) by both lowering the entrance barrier and providing analytic/query tools to make the results of these calculations readily available to the wider observational and experimental Earth science communities.

8. Collaboration/assistance from HPC staff with earth-science researchers at HPC centers.

9. Create a comprehensive reference Earth model that includes both deformation and elastic properties.

Science-Driven Cyberinfrastructure Needs in Solar-Terrestrial Research

Current Challenges to High-Impact, Interdisciplinary Science:

The main challenges identified by workshop participants center around bridging the gaps among the Geospace sub-disciplines, to foster interdisciplinary research.

Challenges in finding / discovering data

- Users do not know how to search for data across multiple repositories, and in general what data sets/resources exist. Data are hard to find, and even harder to transform into the form needed for further analysis.
- Semantic techniques should be available to enable broad discovery and use of data. Tools/libraries that enable the generation of metadata (annotations) in an automated fashion would be preferred.
- Joint data discovery ideally makes use of centralized data repositories or search facilities where all the metadata (and pointers to the data) are queried and made available through a common interface. Complementary to this would be the implementation of a system based on semantic web technologies, which would require that a widely accepted standard vocabulary/ontology (suitable for our community) be put in place that the community agrees to abide to.

- There is a need for encouraging adoption and consistent usage of metadata standards for the essential attributes of both observational and modeling data sets, as well as an agreement on vocabulary to use.
- Getting to a set of "widely accepted standards" is itself a challenge. Also needed are translation tools ("ontology alignment") between different sets of standards, especially where there are already multiple sets of established practices.
- The Geospace disciplines increasingly need better tools for mining our spatiotemporal datasets for features, both known and unknown
- The tools need to be scalable, to work for both large and small datasets.
- Data query: enabling the easy and effective querying of very specific subsets of data in order to tailor the results according to a specific science objective, thus reducing the volume of the data transfer. Good metadata and strong quick-look tools play a big role in this.
- Data volumes are becoming prohibitively large. It is not feasible to co-locate all data sets, or even apply the "old model" of requiring users to download all the datasets of interest onto their own computers to manipulate them locally. Analysis increasingly needs to be co-located with the data, but this is problematic for analysis of multiple datasets, located in different places. Processing and user-driven analysis carried out at these large data centers may provide a solution to this coming problem, but mechanisms need to be in place to allow these providers to develop and support these (potentially costly) capabilities.

Challenges in working with data

- Continuity of data sets (both space and ground-based) over time has an increasing value as our ability to mine and probe these large data collections grows. Ensuring continuity should be a factor in funding decisions. (For example, there are concerns about several older instruments with no successor at the moment.)
- There are similar issues of continuity in the development of data analysis tools as well as instruments.
- Getting the most out of existing or legacy data; ensuring things do not get lost over time as missions or groups end.
- Information about assumptions, sources of error, and methodologies should be included along with the data.
- Need methods to ensure scientific reproducibility by allowing citation of specific data products and processing steps used in a scientific study.
- Need a mechanism for ensuring proper attribution of data sources in publications. It is critical to record provenance of all data to improve future reuse.
- Need better benchmarking/validation of data catalogs for researchers in different disciplines: it is important to have clear quality metrics that allow users to determine which data points are "good" or "bad" for their purposes.
- It is important not to "re-invent the wheel." If someone has "solved" a problem, other communities need to be able to find out about this and make use of it.
- The wide variety of analysis tools and languages in current use inhibits the development of a common set of analysis tools. Clearer documentation and use of software development best practices would help mitigate this confusion.
- There is a need for a strong leadership structure: a project should be run by a single, strong entity with broad community buy-in to ensure coordination.

Challenges in cross-disciplinary science / working with data outside our sub-discipline.

- Data from outside a researcher's field is difficult to find and learn how to analyze.

- An impediment to cross-disciplinary research is that while the same problems might be studied in different sub-disciplines, the observables, scales, and parameter regimes may be quite different.
- It is difficult to find sources of funding for cross-disciplinary research.
- Researchers using data from outside their areas of expertise need trusted catalogs of events and categorizations
- Data integration is needed to enable interfacing and interoperability among diverse datasets.
- Need better support for 'sun-to-mud' efforts. Solutions may be to have more common workshops, and classes offered online by multiple institutions.

Modeling-specific challenges

- It is important to compare and address discrepancies between data and models. Tools are generally not readily available to directly compare model outputs and observations.
- If these tools were available, iteration between modeling and data comparison could take place, allowing ongoing improvement of both.
- While data are often open and analysis code is sometimes open source, the same is not generally true for models (although it should be).
- In terms of modeling: there is a need for better flexibility/modularity in large model design so various groups could "plug and play" their components.

Educational, societal, and public outreach challenges

- There is a dearth of data-science and cyberinfrastructure-related content in the domain-specific academic curricula, impairing the ability of students to incorporate existing tools and best practices into their research.
- Scientists often do not know how to scale up their cyberinfrastructure usage from the desktop to make use of high-performance computing (HPC).
- Students and practicing researchers need training on how to use GPUs and other advanced computing resources.
- Scientists want to share their data in the public domain, but may worry about potential misuse or misinterpretation of the data.

Technical Issues/Challenges

Many of the interdisciplinary science challenges noted above are rooted in technical issues that must be addressed in order to successfully overcome them. The breakout sessions devoted to technical challenges included moderators who are computer scientists, in order to encourage new thinking.

- There is a need to develop computationally efficient capabilities for searching and expressive querying of Large/Diverse/Distributed Data Sets including provenance and data quality. What is of interest to scientists can be very complex to define. With today's high-volume databases, it is increasingly important to locate and download only the portion of data of interest. Propagation delays from one regime to another within the Geospace system make event searches challenging—e.g. how to do correlations to find linked events among data sets with such delays, without downloading all of the data.

- There will be a continuing need to discover, search, and utilize historical datasets, which must be preserved and, if necessary, modernized through metadata indexing to bring them into discoverable form.

- Data providers, especially new and actively maintained services, need to include well-documented APIs (application programming interfaces) and service interfaces, to aid in development of flexible workflows for utilizing the data resources.

- Some metadata standards already exist, but translators/converters are needed for searches bridging solar-terrestrial environments (solar, heliosphere, magnetosphere, ionosphere/upper-atmosphere) to promote interdisciplinary science. Additional efforts to agree on a wider standard of keywords, vocabulary and ontologies would be useful, but difficult.

- A platform and standards for data and software citations need to be further developed and widely adopted. A scheme for searching ranked databases and software according to popularity, usage, and quality would be a useful addition.

- Workshops/tutorials and academic curricula are needed to teach standard tools and techniques for interdisciplinary research to the community (e.g., orbital discovery tool). Community-developed toolkits (e.g. those at SolarSoft, sunpy.org, itk.org) are important sources of cross-platform tools for general analysis. Community involvement in further open-source tool development (e.g. through Github) should be strengthened and encouraged.
- Tools are needed for generalized Event/Object recognition in space and time, and for visualizing multi-dimensional data in large data volumes

Sedimentary Geology

Existing tools, databases, etc. needed for pursuing key science questions:

1. The Paleogeoscience domain workshop previously identified ~140 cyber databases, repositories, or tools, with particular focus on paleobiology, marine sediments, geochronology, and paleoclimate. The Sedimentary Geology workshop added 83 additional cyber resources to the compilation – 38 databases, 17 repositories, and 28 tools. These additions particularly focused on LiDAR, map resources, and tools for use in sedimentary geology and subsurface analysis. Of particular note is that there are very few databases for onshore sedimentary geology, most repositories of subsurface data are state or federal agencies, and the most thorough software tools are commercial.

Desired tools, databases, etc. needed for pursuing key science questions:

2. To forge new ground and develop richer comprehensions of complex problems and systems, sedimentary geology research requires multidisciplinary approaches, easy access to large volumes of geologic and geophysical data, better integration of that data and legacy data, and increasingly sophisticated numerical modeling of sedimentary systems and stratigraphic architecture.

3. A Google Earth-like interface is envisioned with topography, surface, and subsurface geology. The interface would (i) allow a wide range of queries, (ii) compile and visualize a variety of data for different time intervals and geographic locations, and (iii) have the ability to create cross sections from designated line paths and make maps for designated areas and time/depth intervals.

Databases (Geo-referenced; can also be catalog information, not just data)

4. Geologic maps, cross sections, seismic and GPR lines, LiDAR data, macrostratigraphy.

5. Distribution of fossil organisms through space and time (e.g., Paleobiology Database).

6. A better compilation and integration of the available paleoclimatic data.

7. Drill hole that integrates or links across state boundaries and includes locations, formation tops, geophysical logs, cored intervals, core photographs, poro-perm data, thin section imagery, total organic carbon values, thermal data (e.g., vitrinite reflectance).

8. Measured sections of outcrop and core (both referenced by midpoint of section or line in dipping units). Include scanned images of cores, lithologies, sedimentary structures, grain sizes, textures, fabrics, contacts, trace fossils, thin section imagery, poro-perm data, mineralogy and whole-rock geochemical data (e.g., stable isotopes).

9. Sedimentary rock imagery: stratal geometries, sedimentary structures, photomicrographs, etc.

10. Data on age constraints of stratigraphic units, including source and basis of age.

11. Hub for coordinating databases.

Search Capabilities

12. Multi-tiered search engines to access and search different databases.

13. Searchable map-based areas of interest by time, space, stratigraphic unit or topic.

14. Spatial querying for published work.

15. Filtering tools for searching (search engine and tagger).

16. Ability to search by example - an image of the object or a verbal description - and the query system finds things that are similar (fuzzy query for dark data?).

Tools *(must enable range of data formats and conversions)*

17. Template or checklist tool for metadata format.

18. A suite of tools to easily sort/analyze data using available metadata.

19. Ability to map (with contours) all types of quantitative data.

20. A set of tools that will correlate between sections/core.

21. Tools for compilation and correlation of biostratigraphic ranges for different index taxa.

22. Basic sedimentary interpretation tools (e.g., of depositional environment) involving guided questions that direct interpretation process.

23. Capability of determining sediment volumes/thickness/accumulation rates/fluxes from measured sections, logs, seismic data, etc.

24. Open source visualization software for well, seismic, and LiDAR data.

25. Open source visualization software for stratigraphic columns, timescales and other data (biostratigraphic, chronological, geochemical, petrographic, etc.).

26. Higher resolution paleoclimate climate models.

27. Interoperability with the CSDMS (Community Surface Dynamics Modeling System) suite of modeling tools.

28. Ability to track users of particular features to help organize conferences and workshops of people with common interests.

Other

29. Ability to enter the data as it is collected.

30. Continued development of GeoDeepDive techniques - machine learning data-mining to extract info from PDFs and convert it to a database that can be directly queried.

31. Training modules for database creation and entry, search tools, analysis and visualization.

Shaping the Development of EarthCube to Enable Advances in Data Assimilation and Ensemble Prediction

1. Centralized data repositories and services that link existing and future data systems. For example, a centralized community repository could be created for data submission and sharing.

2. Advanced software, tools kits, and services for quality control, in-depth data analysis, visualization, verification, and mining of data (observational and model output). These tools and services need to be user-friendly and accessible by the whole scientific community.

3. Common data formats and frameworks for assimilation, modeling, analysis and visualization.

4. Common data assimilation framework; currently, each assimilation system uses its own framework for data I/O, processing, and running algorithms.

5. Collaboration tools, platforms, and frameworks (e.g., Wiki for data)

6. Server-side processing tools for data processing, analysis, visualization

Structural Geology and Tectonics

1. Workshop participants considered developing conventions and technology for data interchange and documentation to be the highest priority component of cyberinfrastructure needed by the community. This system should be web accessible and allow discovery, access, and reuse Structural Geology data. The scope of such a SG&T Database (or Dataspace) was not developed in detail, although it was recognized that field and microstructural observations would be need to be geospatially referenced. Standards and technology developed by various groups (OpenGeospatial consortium, IUGS Commission for the Management and Application of Geoscience Information (CGI), W3C) were mentioned, and these approaches could be used in the development of such a system.

2. There was agreement that analytical tools routinely used to evaluate structural data should be developed in the context of this SG&T Database. These tools include - but are not limited to - stereonet plotting, shape preferred orientation analysis, rotation of data, calculation of finite strain, vorticity analysis, spatial error analysis, three-point problems, etc.. A specific set of new tools would be focused on processing map data. If convenient and powerful tools were available for compiling and analyzing geologic maps and map data, workers would have a natural incentive to use the tools. At the same time, the map tools could serve as a front end for larger map databases. Maps could be designated as "private" until publication, but once public, they would be available to researchers around the world.

3. A vast amount of Structural Geology data already exists in the form of geologic maps. These maps contain primary data and are at the very core of the field. Most of these are not in digital form, and the workshop participants considered the digitizing of these legacy data to be very important to the community. This task was considered to be a potentially high impact investment in digital conversion. Semi-automatic to rapidly guided digitizing is considered by the group as an appropriately challenging endeavor for EarthCube. The use of cross-sectional data is particularly challenging, because cross sections involve increased interpretation and their vertical orientation is poorly handled in existing map-based approaches.

4. Development of innovative methods to build and visualize interpreted structural histories would be very useful to the structural geology and related geoscience communities.

5. Because Structural Geology and Tectonics relies on integration across the Earth Sciences, scientists and students in this area must use data and tools from other fields in the geological sciences. For example, many structural geologists working in neotectonics need access to GPS and LiDAR data. In practice, it can be difficult to find the appropriate data; when found, the user may not be aware of, or how to use, the appropriate tools to solve their structural problems. At a minimum, maintaining a listing of tools and data is critical. More significant advances would involve cataloging resources for best practices and tool use, in addition to making more accessible interfaces for data from other domains.

6. There was a keen interest in developing digital laboratory/field notebook software for wide adoption to increase efficiency in the field and facilitate data integration. The concept of the science workbook would be to allow a researcher real-time (or pre-loaded) access to all the geological data from a specific region. This science workbook would form the basic cyberinfrastructure for interacting easily and seamlessly with the database noted above. If well designed and made sufficiently adaptable, the software could be tailored in part to be the front end for the structure database; data collected to be immediately uploaded to the structural geology database (although the data might not become publicly available immediately, to allow for field re-checking, etc.). This software would be platform independent and would have to run on devices from smart phones to pads to tablets to desktops. The development of this type of

science workbook would be an important step in developing a cyberinfrastructure for Structural Geology as well as all field-based sciences. Various existing software provides a starting point for defining the functionality and implementation of the science notebook.