User Scenario Summary Matrix. From: Stocks, Karen I.; Schramski, Sam; Virapongse, Arika; Kempler, Lisa (2019). EarthCube User Scenario Collection. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0WQ024C

FILENAME	SCENARIO NAME	DOMAIN KEYWORDS	OVERALL GOALS	MAIN CYBERINFRASTRUCTURE CHALLENGES	DISCOVERY CHALLENGES	TYPES AND FORMATS	SOFTWARE USED	STANDARDS USED
4D visualization_Pavlis_out	4D visualization	3D visualization for geology	To develop and build four-dimensional geologic models that are of use to researchers in diverse communities within the geosciences.	- The big challenge is the software necessary to undertake 4D visualizations to this point. There aren't really any killer applications currently, and all of it has become quite cumbersome--we're kludging together data sources. ArcGIS is designed for application specialists, not for the needs of geologists - Need an environment for data management - An industry/academic partnership to do that? A facility that stores sophisticated maps? A library	Data is on people's computers, not accessible And when you do assemble, resulting data is large Also the detail level of the data can be very variable. USGS isn't addressing the needs	•Data Source ○ Geochronology •Field data •Stratigraphic data, including fossil \•Obscure things that are hard to quantify: x is older than y, in a way that is built into geologic thinking—comparative sources •Data Format ○ Map info that is input into GIS and then adapted into a 3D GIS database •Simple paper map scanned as a raster. •Geochronology sources •Stratigraphic formats from available software •Satellite imagery, including LIDAR formats	Move EarthVision, etc. Paraview	None
Access to Wisconsin Geologic samples collection_Stanley_out	Access to Wisconsin Geologic Samples Collection	Informatics; Geochemistry; Geochronology	To have a portal to the data and samples in the Wisconsin Geological and Natural History Survey (WGNHS) repository to make them more discoverable.	- need to standardize the collection. Some samples have IGSNs, but not all. Legacy samples don't always have the metadata necessary for an IGSN. would like a more efficient way to get IGSNs - no clear place in the workflow where IDs and metadata are captured. Time consuming and effort is limited particularly for legacy sample work - need to build a portal	- They need to build a portal to make their sample information discoverable; but she says this is not hard- there are lots of good exemples out there - but is workd and resources are limited. - it would be helpful if there were an easy way for their database to communicate directly with SESAR (the ISGN registry) - need to develop metadata templates for uniform data collection.	- Data Types: Data and metadata about geological samples: geospatial, geochemical, descriptions of data, images, scans, geophysical data, and digital elevation models. Data of the different samples are tied together by depth (i.e., in the earth). - Data Sources: field, chemical analyses (XRF devices), digital elevation model (DEM) output - Data repositories: SESAR IGSNs repository. Project hosts their own data on WGNHS servers. - Data Formats: Excel .xls, Arcmap shape files, layer files (e.g. lidar dataset), jpeg, nef, .kmz - Data Volume: 10.3 TB - Data Velocity, Variability: not stated as challenges - Data Veracity/Quality: needs to be addressed. Interviewee is working on a workflow for quality. - Data Variety: many different kinds of data need to be brought together.	ArcMap. ArcMap, Petrel (can view samples through Access, but it's in Microsoft SQL server), Adobe Photoshop and Illustrator, Microsoft Office, LabelMatrix They are currently deciding on which descriptor software to use.	- USGS metadata repository – National Digital Catalog. - NCGMP09 – National Geologic Map Database Editorial note: these are listed in the standards section, but it is not clear if they are standards, or data repositories/resources.
Aeolian processes field work_Martin_out	Aeolian process field work data management		To understand the stress of wind on a barren sandy land surface in a field context (active aeolian processes). In other words, what is the amount of sand moved by the wind and turbulence?	1) no storage location or scheme for his data, no agreement in the community 2) data and software in 2 different places (not clear that's a big challenge for him) 3) lack of measurement accuracy combined with need to invent conversion algorithms could lead to unreliable data	Needs better ways to share data with collaborators Github doesn't allow storage of large data file Google Drive stores large files but doesn't enable SVC and development collaboration Hard to know best way to document data for reuse by other scientists No precedents or human guidance in field of Aeolian process for how to store and manage data No existing systems for storing and managing Aeolian data Different terminology in Geomorphology vs. Atmospheric Wants credit for the datasets even though he wants to share them Tracking use of datasets (DOIs?)	- Data Source: field data, including instruments, sample collections, and qualitative observations - Data Format: sensor outputs, such as ASCII comma delimited files; field notebook- scanned as *.pdf; spreadsheets as *.xls; photos as *.jpg - Volume (size): 22 GB of raw data - Velocity: Batch. Data are collected during 1-day or few day intervals over a period of a few weeks. Data is processed after all data have been collected. - Variety: Data from the sensors consist of pulse counts, voltages, serial (RS-232) outputs, and amperages. - Variability: temporal scope is the same for all observations (e.g., few days at a time or during one afternoon). Temporal frequency of 50 Hz (anemometers) to weather station every minute; data are collected from sand traps every hour (although interval may change depending on conditions). - Veracity/Data Quality (accuracy, precision): They trust the manufacturers of the sensors regarding calibration of instruments. Quality of physical measurements (e.g., sand traps) is based on standards developed by previous studies. - Data Types: sensors records, sand sample analyses, photos Identify sensors, portals, and final and intermediate data products that your project generates as outputs. Intermediate data output products - *.mat files (MATLAB).	Proprietary for data logging MATLAB for analyses Excel for spreadsheets	None

Anion and cation concentrations in rivers_Peucker_Ehrenbrink_out	Major dissolved anion (SO 4 , Cl, Br, HCO 3) and cation (Ca, Mg, Na, K) concentrations in rivers outside the tidal influence	Biogeosciences; Geochemistry; Global Change; Hydrology; Oceanography: chemical	Understand state parameters of carbon dioxide uptake by mineral weathering reactions in drainage basins in the North America.	- data repositories can't be searched in all the ways desired - some data are in publications/reports/grey literature and not searchable databases - some data are "dark" (not online) - Not having metadata and formally defined semantics for key terms results in less efficient resource discovery. - datasets are discrete and sparse - relevant data may be taken on very different temporal scales (once per yr sample analysis vs every 15 min from in-situ sensor - data not in usable formats need transformations	If a user could sit on her/his computer and link to a search engine, be able to type in the name of a river in North America or select a geographic area and subsequently be able to pull up data on major dissolved anions and cations, and do so in a simple fashion, I would consider that a measure of success. Specific searches desired: - Critical issue is dealing with the tidal zone. The farther the sample location is towards the ocean, but outside the tidal zone, the bigger the drainage area is that the data represent. Must know relation of data location to tidal zone. - Query system may correctly interpret concepts such as “major dissolved anions” or “tidal influence”, or “river”, and map them to O&M. - GML constructs such as “result”, “feature class”, “sampled feature”. - Expands the term “major dissolved anions” to a set of individual anions, concentrations of which are actually contained in the sources. - Then we have to define the chemical constituents themselves, major/major which have chemical names—define and utilize them effectively. - The same goes for units: the really necessary thing is getting access to the data.	- Data Sources: in-situ sensors, field data, sample data, lab analyses - Data Types: isotope ratios, elemental concentrations, water discharge, supporting methods information. Tidal zone location. - Data Formats: csv, asci, txt - Data repositories: Gauging stations (USGS and Environment Canada - Water Survey), Web-accessible geochemical data (USGS, Wateroffice Canada, EPA, CUAHSI-HIS), udigitized libraries/collections. - Data Volume: not a challenge. - Data Velocity: sensor data can be every 2 min. Integrating data from different velocities is a challenge, but the overall velocity is not. - Data Vearacity/Quality: requires sufficient metadata to assess.	- Natural language textual analysis - CUAHSI (Consortium of Universities for the Advancement of Hydrologic Science, Inc.) - LoadEst and LoadRunner (to estimate daily, monthly or annual fluxes) - ArcGIS or similar GIS products.	- Open Geospatial Consortium’s (OGC) Observation and Measurement (O&M, also ISO 19156; Cox, 2010); - CUAHSI’s data standards; WFS (Open Geospatial Consortium Web Feature Service Interface Standard); - OGC Geography Markup Language (GML, also ISO19136; Lake et al., 2004; Portele, 2007) general feature model; - SESAR’s IGSN (provided water samples have been registered)
BioGeoSIGiPlan_Goff_Thessen_out	Predicting best crop location	Global change; public issues	Increase agricultural production by identifying which corn genes and variants are responsible for allowing corn to reach yield under different environmental conditions; allow seed producers to predict which varieties will do well in areas based on climate model predictions. This requires the analysis of genetic data with yeild information and meteorological data, soil data and other environmental data.	- Data are distributed among multiple sources; some data is proprietary, 49formats are heterogeneous and need manual effort to integrate 49 analysis requires demanding computations only possible on the largest systems	1) Average corn yields still fall significantly short of maximum yield potential defined by genetics in ideal growing conditions. 2) Data has to be brought together using brute force and manual methods: a) All of the heterogeneous data sets from the distributed sources have to be re-entered in a new software environment and integrated around space and time. b) Significant time has to be taken for downloading and reformatting—differing time scales. 3) Weather changes all the time and corn is in the ground for several months: simple averages won’t work. 4) Most traits of interest that have an impact on yield are complex traits, i.e. traits controlled by many different genes. So computational analysis needed to identify the genes underlying complex traits is very demanding and only feasible on the largest systems currently available unless new analysis routines become available.	Data formats: pdf (tables and maps), Excel, FASTA (genetic), GRIB (climate projections), netCDF, CSV, JSON, MS Access. Data Types & Sources: corn seed genetic variation data (companies and putlic sector), climate projections (IPCC), crops location data with productivity by state and county (USDA), meteorological data (NOAA NWS), soil data (USDA Natural Resources Conservation Science). Data Volume: "Big"	Current: R, Genome Workbench, BioPython (for genetic data), netCDF library, MatLab Prospective: GIS	NOAA, USDA, iPlant, and seed companies have standards. The hope is create a new integration and new standard.
Cenozoic cetacean diversity_Uhen_out	Cenozoic cetacean diversity	Biogeosciences; paleoceanography	To compile all existing data on cetacean diversity through the Cenozoic, in order to understand what drives diversity patterns.	Potential solution by EarthCube: Time and location is what links everything together, and can be used to link datasets together. Every sample has coordinates and can be traced to a time period. If EarthCube could help with developing some kind of database to help resolve the chronologic challenges in his work (e.g. developing a geologic chronologic database.) - Sustainability - time investment required to enter the data, data cleaning	He used a crowd-sourced database - Paleobiology -- for paleontology to manage and organize the data. 387 people have contributed. DD challenges - Maintenance model - time investment required to enter the data, data cleaning (does GeoDeepDive offer any lessons?); completeness, sustainability, accuracy	*.csv Just working with text documents.	R Excel Adobe Illustrator	None
Chemical properties of surface and groundwater_Lunch_out	Chemical properties of surface and groundwater	Biogeosciences; hydrology; global change; informatics	Facilitate diverse ecological forecasting through a better data management system for water/groundwater chemical analyses that can streamline data uploads, relate different data taken at the same site/time, and better user access.	- diverse data, many data providers, diverse users. Relationships between related data not currently captured. - Data are not currently all centralized. - need better QC associated with ingest.	Most of the instrumented data, like towers measuring atmospheric variables and the like are data set in time series with a time stamp and a value, time stamp and a value, etc. You can have millions of these time series for each instrument. From a database point of view this is relatively simple. When you start looking at the observational data, however, it is very different: the data isn’t simply a time stamp and a value anymore. You might have context, such as “I measured water chemistry on this water sample X,” and “I also measured these other products on this other sample.” So it gets more complicated. We need that linkage built into to system, such that a sample taken will ultimately build toward a database that has all the interrelations and is more complicated than just one simple product.	- Data Source/Type: analyses of water samples, with sample lat/lon, date, and elevation, etc. - Data Repositories: NEON data portal - Data Format: mainly CSV on collection side. Oracle back end. EML for metadata - Data Volume: not an issue. Largest is multiple MG per year for all the sites. - Data Veracity/Quality: is a concern. Writing down sample data incorrectly is an issue. Can be date wrong, sorted wrong, ented twice, etc.	Oracle database back end. Liferay software.	Ecological Metadata Language (EML) for metadata

CHORDS_Kerkez_out	CHORDS (Cloud-Hosted Real-time Data Services for the Geosciences) – access of real time data	Hydrology; Natural Hazards	Create a cloud-based infrastructure for real-time geosciences data needed for improved flood forcasts and other applications. Provide the best possible estimate of what is happening on the ground at any time and then use it to forecast into future.	- real-time data acquisition & management - Scaling the architecture up to the number of sensors		- Data Types & Sources: real time field sensor point data (weather sensors). Radar (model output as product/future goal.) - Data repositories: "a number of repositories" including for radar data - Data Formats: csv, NetCDF, SensorML, "rudimentary" database - Data Volumes: could be huge. High resolution radar, many streams - Data Velocity: real time or near real time - Data Veracity/Quality: need QA/QC on real-time data	CHORDS visualization SEEK JavaScript Online script with geospatial software that supports raster and general web site frameworks	SensorML NetCDF
Climate feedback of mesoscale cloud-field organization_Kuo_Nair_out	Climate feedback of Mesoscale Cloud-Field Organization Using an Event-Based Approach	Informatics. AGU poster on the topic was listed in "Public Affairs"	The scientific objectives of the use case are: - To examine the evolution (i.e. lifecycle) of cloud-field organization of individual mesoscale arcs, trade wind cumuli, and tropical cloud/storm clusters. - To quantify and characterize their evolutions, - To accumulate long-term (i.e. decadal) statistics of the evolutions, - To identify changes in their long-term characteristics, - To ascertain the presence/absence of correlation between their changes, and - To discover the feedback mechanism(s) of their changes, if there is correlation.	Most earth scientists are research oriented and phenomenon focused. For most events, there are few long-term records that exists. An exception is hurricanes or El Nino events. - Currently, scientists move data from their local computer or storage resources before they start their analysis. This process is becoming too unsustainable because the data volume is too large. Also, there are often multiple copies on different computers (e.g., redundancy). [reducing the need to download data was highlighted as a major goal] - There are lots of different kinds of data being generated. E.g., Radiometers - different wavelengths; LiDAR radar - sends out signal; Polarimeters; Synthetic aperture radar, etc. How can a scientist deal with so much variety of data? - Model simulations are now conducted at higher resolution, which requires more velocity. This has implications on applications. For example, if there is a disaster, data must be analyzed more quickly for decision-making. - How can we be sure that we are collecting and analyzing data accurately? Reproducibility is also needed.	Investigators queries Para-DIAME to find all of the datasets associated with cloud formation and potential drivers (e.g, land use, and atmospheric characteristics), as well as in-situ and remote sensing observation, in the region of interest.	- Data Types Sources: Radiometers - different wavelengths; LiDAR radar - sends out signal; Polarimeters; Synthetic aperture radar, etc. Long-term data on cloud formation and evolution; processes, e.g. aerosol and land use, that could impact cloud formation and evolution. - Data Repositories: Para-DIAME (is a data and compute environment in the cloud) - Data Velocity: high resoluton models sometimes need to be analyzed quickly in cases of disasters - Data Veracity/Quality: is a challenge - Data Variety: is a challenge due to the many data types	- Para-DIAME (is a proposed data and compute environment in the cloud). SciDB: A parallel distributed database management system based on array data model. - Spark (on top some supported storage backend) - Hadoop (MapReduce) on Hadoop File System, HDFS	
CO2 repeat hydrography data_Kozyr_out	Data management for ocean carbon dioxide measurements: repeated section hydrography	Oceanography	To manage the data that is developed from CO2 measurements in seawater: total carbon dioxide, dissolved inorganic carbon, salinity, partial pressure, and pH. In order to: - Calculate anthropogenic CO2 in the ocean; scientists use this to separate manmade changes throughout a given year - Determine how much is sunk into the ocean. Calculations made give us sense of ocean absorption. - Repeat sections determine how much these measurements change, for example pH level	- Uniting biological and hydrographic data. This needs integration/interoperability among data centers, and also community guidance. - Data quality is a challenge: the quality of measurements can vary among groups collecting them. Some data needs to be thrown out. He frames this as a funding problem, not a CI problem. - Metadata is/was not sufficient to know critical factors about how a water sample was taken and analyzed.	Not detailed but could perhaps be extracted for the biological/hydrographic integration he suggests.	- Data Sources: chemical measurements from in-situ sensors and analyzed water samples collected from ships and moorings/buoys. - Data Repositories: Mercury, Web-Accessible-Visualization System (WAVS), CDIAC - Data Formats: CSV or TSV spreasheets coming in, CSV and NetCDF for publishing out. Images, simulated sky maps. XML. - Data volume: 100's of MB to GBs - Data Velocity: not clear if he is working with real-time data, or delayed mode.	WAVS data synthesizes data automatically and produces visualization relationships between databases (developed within project) Ocean Data View - for data manipulation and QC Excel CDIAC program for CO2 calculation app	DOIs for data. Mercury metadata.
CPO and 3D strain data integration_Mookerjee_out	CPO and 3D strain data integration	Structural geology; tectonophysics	More efficient and complete generation of information from existing crystallographic texture data by improved community data management and sharing mechanisms, and by developing statistical models that integrate these data with three-dimensional strain data. With the ultimate goal of understanding kinematics of faults and shear zones, so that we have a greater understanding of how the structural changes inside the earth on large-time scales can be made, for instance, understanding more completely how mountains are formed.	- overall: good models need the integration of data from many studies and researchers. Data exist (crustallographic and 3D strain data), but are stored in personal collections. - Need a place to store data and analysis scripts. - Need to motivate community to upload/share data - Need middleware to translate data formats: Individual researchers are developing their own conversion scripts, which duplicates effort. Some formats are proprietary. - Datasets can be large - up to 1TB each - which is too large for many researchers to share on their lab websites or download for local processing - Need better analytic software for meta-analyses of these data - a lot of analyses are still being done by hand, inefficiently. - Need standards and best practices for sharing these data	- wants to online access to crystallographic and 3D strain data, as well as converstion scripts. Much not online now. Did not specify search/discovery requirements, but could probably provide that information if asked.	- Data Types: crystallographic fabric/texture; 3D strains; chemistry data (ESX), EDX and EBSD data. - Data Formats: many. .txt. Oxford Instruments uses a *.cpr for EBSD data. Many in proprietary formats. - Data Repository: Strabo (in process) - Volume: range from <1MB to 1TB per sample. Electron Backscatter Diffraction (EBSD) produces such large datasets that they are not easily shared or even stored on an individual researchers' webpage - Data Variability, Veracity/Quality, Velocity not named as issuesl	- Matlab, Mathematica scripts - Middleware for translating between data formats (particularly between data files generated by different analytical equipment and different equipment companies) - Analytical software that allows for analysis of large, integrated crystallographic and 3D datasets (existing tools for large scale analysis are not sufficient.)	None that he is aware of
Deep search of scientific data_Mattmann_out	Deep search of scientific data	Informatics	To make web data that is behind web forms, logins and the like more searchable, particularly data in HTML, ASCII files, and array-based formats like NetCDF,.	Much of the data on the web is not searchable.	See main challenge: Much of the data on the web is not searchable.	ASCII, Excel, HTML, NetCDF	DARPA Memex	The standards are those developed by previous projects using Nutch and Memex

Deformation of active volcanoes_Freymueller_out	Deformation of active volcanoes	Geodesy and gravity	To understand how, where, and when magma accumulates underneath active volcanoes, and what is involved in the eruption process for getting magma up the surface.	- latency in data collection of weeks - Quality control of data - Having accurate and up-to-date metadata from multiple contributors - Usability of InSAR data	See CI Challenges	Formats: RINIX – for raw data; time series of positions for each site are stored in a simple ASCII file with fields separated by whitespace Data types: GPS and InSAR	MATLAB JPL - GIPSY/OASIS C, Fortran (inhouse)	Processed by UNAVCO to turn into metadata
Digital rocks portal_Prodanovic_out	Digital Rocks Portal	Geology	Develop an open source and easy-to-use repository of 3D rock images (called the Digital Rocks Portal – DRP).	- Sustainability of data/images, including web pages and software to communicate with users - Usability/easy browsing - Proper image display - Lack of standards - Many different formats used	- Usability/easy browsing - Proper image display - Lack of standards	- Many formats used. They prefer TIFF	ImageJ Paraview Django Dropbox/UTBox Bitbucket	Author and data description standards on datacite.org Ones associated with DOIs
EarthCollab data center_Mayernik_out	EarthCollab data center use cases	Polar programs; Earth Science; CISE; AGS; and OCE	1) Developing of a novel semantic web cyberinfrastructure using Vivo software; and 2) Developing semantic web models specific to the needs of earth sciences by using the data center use cases of UNAVCO's geodesy research and a Bering Sea project to provide the technical details and domain science content for the models.	- Challenge: Getting the links established between components in the semantic web, e.g. between publications and datasets, is time consuming manual process now but could be done with software. This is more of a human problem than a technology problem, as standards are just starting to be developed to help guide people's documentation of information. There are many people working on trying to solve this problem. Some solutions underway in the larger community: 1) Assigning DOI's to projects and project components (e.g., datasets), and 2) Create standards among publishers so that these connections and tracking capabilities happen at the point of publication. Efforts made in their project: They are using a combination of search engines for data mining and manual searching. Manual searching is working fine for their small project (on the order of thousands of data entries), but it is not a scalable solution. As their project grows, this is likely to become a problem. - Challenge: Data variety. Without available standards, each data center/institute uses different formats. This data variety is a challenge because all must be converted to RDF, so that they can be used by Vivo. For example, they must map time ranges into an RDF structure, and this is time-consuming. It would be nice if this process was more efficient. This is a challenge that all semantic web projects face. On the other hand, the diversity in the data formats is also beneficial. If you try to standardize everything, it takes away the value that diversity also has to offer (e.g., more opportunities for innovation). Efforts: Currently, they do all of this translation work manually. They are writing translation tools themselves with the help of software engineers.			Vivo	RDF Ontology - VIVO-ISF, GCIS, DCAT
Environmental Seismology_Hsu_out	Environmental seismology	Geomorphology; earth surface processes	To use broadband seismic networks and other seismic data to learn about environmental processes that create vibrations on the surface of the Earth (not earthquakes). For example, to understand sediment transport in rivers using seismic data. To better integrate seismic and non-seismic data for diverse science (geomorphology, ecology, etc.)	- Seismic data is held in IRIS, but non-seismic data are not and are difficult to discover. Making those data accessible has the usual set of challenges. - Need the architecture to be able to search existing infrastructure and existing databases so that the discoverability of the relevant data is improved.	1. Determine the data and data types needed to solve the question: For example: how much sediment is being transported in rivers? Seismic data can be used to address this question since the process of sediment transport creates vibrations on the ground. 2.You need broadband seismic data, river gauge data, river discharge data, precipitation data 3. One method is to query and see where these different data overlap: I want to know where there is a seismic station within 2km of a large river: large river defined by X discharge 4. Search all the different data accessible and try to answer where these data overlap 5. It's probably in some raw or not very processed state. The next step would be to set parameters to get derived products that could be compared. 6. Then look for correlations and analyze them to figure out some result values that can tell you about the question more broadly. 7. Reiteration of those steps depending on subsequent results; in an ideal workflow it could be re-run multiple times.	- Data Source/Type: 1) Seismic data in IRIS and smaller investigators on short term projects that aren't in IRIS. 2) Non-seismic data: river discharge/gauge that is luckily well-organized, GIS files, imagery from satellites, photos, (landslides) informal documents and emails about events, suspended sediment and bedload sediment measurements (not spatially extensive as seismic network) - Data Repositories: IRIS seismology data center - Data Format: IRIS format, maybe SEED data, ArcGIS/QGIS files, txt, doc, photos in tif and jpg, csv. - Data Volume: Seismic data is GBs, non-seismic is MBs or larger - Data Vareity: high due to the different types/sources, and also level of processing (raw, derived, mashup) - Data Velocity: is continuous for seismic data. - Data Veracity/Quality: the database (assumed to mean IRIS) does quality contro.	homegrown tools in MATLAB and Python. IRIS provides software. SQL for databases. ArcGIS, Excel are essential tools.	IRIS has some standards, but for the Earth surface process data, there are no larger standards (and this is a challenge).

Experimental debris flow data_Hsu_out	Experimental debris flow data	Geomorphology; earth surface processes	To facilitate and make more efficient the modeling of erosion/weathering by developing community repositories for data and scripts, with examples standards for the data and examples of how to document workflows.	- There is no community repository that will take all their data or their scripts. Having better access to data and scripts would make research more efficient - she was rewriting code, as a non-programmer, that had likely already been written by someone. - The field has limited access to software technicians who have sufficient knowledge of their field to write, and document for reuse, needed software. Ontosoft exists, but her community has not yet bought in - their data volumes are large - too large for some repositories. - Insufficient community standards for the data she collects - NSF does not fund long term repositories		- Data Source: laboratory experiments - Data Types: Force, video, topography (from laser and from still images. - Data repository: SEAD - Data Formats: MATLAB, .avi for videos - Data Volume: <1 TB for her data. But in aggregate, data volume in community is large enough to be problemantic for repository - Data velocity: not relevant - Data Veracity/Quality: instrument specific	Software used: Matlab software, Flume specific software designed by academic research group. Custom Visual Basic program to monitor sensors in real time.	community standards do not exist and are needed
Experimental sedimentology data_Straub_out	Storage and use of experimental sedimentology data	Earth Surface Processes; Stratigraphy; Hydrology; Marine Geology	Understanding how stratigraphic surfaces are related to changes in geomorphic surfaces that existed at certain times on the Earth's surface through both experiments and modeling.	- My data is currently stored in diverse locations for different people in my community and in different databases. They have been mostly good about allowing me to utilize different places because I'm considered a test case. - One of the problems linking experiments to numerical models is that often the numbers that exist between field and numerical models versus those collected via experimentation are challenging to deal and there is not standard or uniform means of combining this data. - There is a challenge to see if our models are producing similar stratigraphic models after multiple runs but we don't have the temporal depth to be able to measure all these model runs over a long period of time. - Funding and reliability of funding for CI is uncertain. - PIs need to understand what they should be doing with data, and this "needs to come from NSF, a roadmap, because we have enough challenges as it is. "	- the end goal is to connect expiremental data and models, but discovery vision not given.	Data Repositories: SEAD (Sustainable Environment through Actionable Data), GeoPRISMS (Geodynamic Processes at Rifting and Subducting Margins), SEN (Sediment Experimentalists Network), his personal holdings (and presumably those of similar researchers) Data Types: Numerical models; LIDAR and topographic maps; seismic data; experiments in lab and field sediment data. Data Formats: DEMs of topography ar ASCII xy, SG for subsurface stratigraphy and others, binary data formats - often proprietary, header information about grain size distribution Data Volume: Any one map could be 100mbs, and we're collecting 100 of them, so then you have large masses that include sizes less than 1mb up to 100mbs Data Veracity/quality: Some of the issues that we have with respect to data quality are topographic measurements that are well below 1mm in accuracy. We have challenges with the LIDAR data and how quickly it can be collected at higher spatial resolutions, but you want them to run in a reasonable format and how often you are going to run them is important. There are similar problems when measuring using instrumentation, including Doppler measurements: how accurately you can relate these is a chronic challenge.	Software: Primarily our work falls between MATLAB at the field-scale and seismic interpretation packages developed by the energy industry. These include the industry standard Kingdom Suite. Another program used is called PETREL. Excel for visualization, of course. R and Python, other than MATLAB and PETREL, are the two other significant programs we use in our lab. - CSDMS is a source of models.	We have our own best model for our data, which we have carried over to SEAD.
Field mapping and collecting structural field data_Whitmeyer_out	Field mapping and collecting structural field data	Tectonophysics; Education; Structural geology	Develop a 4D framework for holding diverse geospatial data from field areas (stratigraphic, paleontologic, geochronological, etc) and an app for downloading a section of data of interest for maps to take into the field and upload new data. The ultimate goal is to build 4D tectonic models of geologic evolution	- a diversity of data relevant to 4D tectonic models exists, but is not easily integrated with each other spatially (into maps), or connected to models. The goal is maps with a historical (time) context. Collect point, line, and polygon data in the field into a data structure that would have fields for associated information. This information will include lithology, structural orientations, stratigraphic, metamorphic, or igneous features, and fields for notes and locations, data and time stamp (color-coded) - the data span multiple scales from thin sections through outcrops to regions. - no good app exists for customized downloads and views of data, and then collecting data in the field that integrating it back into the main database. - lots of components exist, but have not been put together. - some important data in publications/papers.	- having diverse data interoperate through 3D maps layered with Digital Elevation Models. Space is the common element.	- Data Types/sources: legacy maps and field outcrop maps, stratigraphic data/columns, paleoenvironmental, paleontologic/fossil, geochronological, Digital Elevation Models/topography, seismic/subsurface, LiDAR and photogrammetry as point clouds, 3D remote sensing, lithological point data, geodynamic models. Historic and contemporary. We also use outcrop, hand sample, and thin section annotated photographs. So field data, model data, lab data. - Data Repositories: Paleo DB or "fossil database" - Data Formats: mostly geospatial formats like shapefiles (ArcGIS, QGIS), KML, but a variety of data types mentioned - Data Volume:	- ArcGIS, iGIS in iPad in field, 3D COLLADA models, Google Earth. GigaPans. Structure from Motion (SFM) photogrammetry software for 3D analysis. - For field data collection: Field Assets, StratLogger, and various photo sketching apps. I have also used GeoFieldbook, but it needs to be updated for iOS 9. - STRABO is a good system in development, but it is not sufficient: not customizable for field data collection.	OGC standards for 3D modeling
Finding geochemical data and sample metadata_Carter_out	Finding geochemical data and sample metadata to complement investigators' data	Geochemistry; Solid Earth Geochemistry	Compare new major, trace element, and isotopic data from the Arctic to the composition of mid-ocean ridge basalts (MORB) from around the world, by connecting data held in multiple repositories.	In general: relevant data are held in multiple repositories, cannot be queried across in an integrated way, data are not connected across the repositories (e.g. by IGSN), and cannot be queried on all the aspects desired.	- Overall, want to discover and access :"A collection of data from PetDB and if possible external datasets, citations for datasets, documents, or related publications, along with any metadata, images, citations, etc. that exist in external resources such as SESAR." - Want to select data from a particular analytic methods (ICP-MS inductively coupled plasma mass spectrometry, and not from LA-ICP-MS Laser Ablation Inductively Coupled Plasma Mass Spectrometry). - wants to find high quality data from specific rare earth elements at specific tectonic features called spreading centers.	- Data Sources: field data, sample data, lab analyses of samples - Data Types: raw, lab data, pre-digitized and recorded. - Data Formats: not specified. - Data Repositories: SESAR, PetDB, EarthChem, "other external databases and online documents, metadata, images, and citations" - Data Veracity/Quality: the need to find high quality data was stated, but no indication of whether she has the info she needs to make that determination - Data velocity: discrete data (not mentioned as a challenge) - Data Volume, Variability not mentioned as challenges	None mentioned	- IGSN for samples (International Geo Sample Number) - repository specific data/metadata formats/content

GEO-inspired Data-Enabled Materials in Aid of Society_Dera_out	GEO-inspired Data-Enabled Materials in Aid of Society (GEO-DE MAS)	Mineral physics	to understand chemical and physical properties and behavior of minerals for better disaster prediction (volcano eruptions, earthquakes). On the technological side, to develop better materials for engineering, electronics, energy storage, etc. by learning from minerals and other geo-materials.	-Need more funding and more support for putting the entries into the database to keep up with competition. Currently 50,000 entries. - Want databases that are accessible. The databases should be more comprehensive and cover the right disciplines (seismology, for example) - Need training on experiments, how to run simulations - Need more efforts, people to build good, stable software for the community	-Databases aren't comprehensive, accessible	•Data Sources: - various forms of digital imagery, including x-ray diffraction and fluorescence, as well as measurements of thermodynamic properties that include e.g. melting temperatures, thermal expansion, etc. The synchrotron facilities provide much of this data. - Density functional theory calculations •Data Formats: - The x-ray images are stored in standard graphics format (e. g. tiff), as in proprietary compressed image format specific to detector manufacturer. - Crystallographic information is stored in standardized Crystallographic Information File (CIF) format, established by the International Union of Crystallography (IUCr). There is a commission that oversees the definition of these standards and keeps them up to date. - Some of the data (e.g. Raman spectra, integrated powder diffraction patterns) are stored as simple ASCII files.	Excel IDL MATLAB Open-source, free s/w developed at universities for analysis of crystallographic experimental data	Unclear
GeoLink semantics and linked data for the geosciences_Arko_out	GeoLink: semantics and linked data for the geosciences	Marine biochemistry, ocean sciences	The GeoLink Building Block "Leveraging Semantics and Linked Data for Data Sharing and Discovery in the Geosciences" seeks to better connect distributed resources already available from existing infrastructure to support more efficient geoscience research. GeoLink aims to provide improved discovery of and access to data by leveraging Semantic Web technologies and layering those additional technologies on top of existing data infrastructure. - understanding distribution of trace elements and isotopes in the global oceans	- Missing persistent IDs (by cruise, by dataset, and more) - Problems with updating data already in			Chemical analysis Excel PhP HTML OWL	RDF
Information tracking for deep-time project_Ritterbush_out	Information tracking for a large collaborative deep-time project	Paleontology; geochemistry; ecological modeling	To create a community data portal that integrates paleontological and geochemical data about samples (rocks, photos, microscope slides) for use by those sciences plus ecological modelers. Could support important hypothesis testing.	- no system holds all the information about a sample and what has happened to it, plus the underlying stratigraphic framework. Very time consuming to try to capture information manually, data are all split up in "weird" repositories. - need to accommodate age models with uncertainty/confidence level in the dating of samples - GIS works for the spatial component, but not the temporal component	Be able to find all the information about a sample, whether paleontological or geochemical	Data Types: information about samples (field and lab), including microscope slides and images of slides; geochemical data from samples, age models, stratigraphic data Formats: GIS, Excel, Pivot Pilot (not clear exact formats) Repositories: multiple, not named	Excel Pivot Pilot Adobe Illustrator Possibly GIS	none mentioned
Inland water communities and water chemistry_Ehrenbrink_out	Integrating the inland-waters geochemistry, biogeochemistry and fluvial sedimentology communities	Geochemistry; hydrology	To understand the capacity of rivers to transfer chemical signals into the ocean. This is about the interaction about inland waterways and oceans, generally, mixing of chemicals, pollutants tranfer, etc.	Data not reliably compiled, except that PANGAEA has taken it on (EMEA). Plus USGS info is not (easily) accessible to non-USGS affiliates (Scribe: Why not?)	See CI challenges	UTF-8 Unicode		UTF-8 Unicode
Intermittent Aeolian Transport on Earth and Mars_Swann.docx	Intermittent aeolian transport on Earth and Mars	Planetary Sciences: Solid Surface Planets	To use lab exeriments to observe and quantify the threshold and characterize the modes of movement of aeolian sand transport on Earth and Mars; to improve models of aeolian transport.	- field data from two instruments must be manually integrated. Having one acquisition system would help. - she collects data on her laptop, and it fills up. Would be better if data were collected on a computer with more space. It also takes time and money to make copies of the data for backup. - Primary challenge: there is no repository for video and flow velocity data. She gets comparison data out of publications, which limits the field's progress. - insufficient experimental protocols and standards exist.	There is no repository where she can find comparative data - she is searching for data in literature.	- Data Sources: Wind tunnel experiments. - Data Formats: *.dat that gets changed to *.txt for Matlab, and then becomes *.mat. Images are *.avi. Movies are *.mp4 and *.mov. - Data Volume: She saves most data at *.dat (reduced size), so she has lots of data in files of about 40-45 MB. Data volume is a challenge. - Data Velocity: They generate about 7GB + 40-80 MB, per day. - Data Variety: different instruments provide different formats.	Software: VLC to splice videos, Matlab for analysis. They take *.mpeg from VLC and bring it into Matlab to extract the frames.	None mentioned. Lack of experimental protocols is a challenge

Isotopic composition of precipitation_Suarez_out	Isotopic composition of precipitation		To undertake paleoclimate studies over time, such that you one is able to acquire information about precipitation rate, evaporation flux, etc.. Long terms trends are the overarching drivers for the paleoclimate.	- Navigating the data sites can be a nightmare - Sites should be search friendly and navigable	- Problems using IAEA data - A plug in for Google earth doesn't work outside North America - Get data from other random sources - Wants a public site that has everything - The biggest challenge may be whether or not the data exists in certain locations to begin with. There may not be data sufficient for precipitation on top of Mt. Everest as one extreme. It's possible this could all be complicated by the fact that some of what we're asking for hasn't been recorded yet. - Resolution of the data - it varies	- Mostly from Excel	Excel ISODAT PowerPoint Google earth Origin (for similar work)	None
Lake Tahoe as an example of a large inland water body study--Geoffrey Schladow	Lake Tahoe as an example of a large inland water body study	Limnology	To combine real-time and actively sampled data to understand how physical processes in Lake Tahoe, over both short term and long terms, are impacted by ecological change in the area.	- have 3-4 home-grown database systems for their multiple data types that take effort to develop, aren't compatible with each other, and are problematic when someone retires or budgets fluctuate. Don't have the money for Oracle, don't have CS expertise to build better, and their scientists are not well-trained in this. - There is a need for real-time quality controlled data, But the QC processes are time consuming, so this can't be met. Would like automated, artificial intelligence approach to real-time validations. "this may be the biggest ongoing dilemma." - Having one data access architecture serve different user types with different needs/expectations - can't calibrate most instruments remotely - some data needs to be collected in-person	- serving data to diverse users (public wants downsampled, science wants high temporal resolution, etc) using home-grown systems developed with little CS expertise.	- Data Types: ecological and water quality data station, processed image data, simulation data, sequence data. Imported LIDAR data. - Data Sources: in situ sensors (field data) - streaming and manual readings. Model and analytic outputs. Molecular sequencing - Data Repositories: "NOAA data sources" - Data Formats: CSV (put into SQL database): essentially GPS location, timestamp, and a set of measured parameters. - DAta Volume: "Huge": 10MB per day for water quality. 20TB/day for LSST (but not clear if they are archiving this or using it)	Software: We utilize geospatial software, MATLAB, and a variety of statistical packages. We use various statistical and database programs where possible as well. - SQL for database	- "there are not standards to the data types we collect or store."
Land use monitoring with unmanned aircraft_Wyngaard_Barbieri_out	Land use monitoring with Unmanned Aircraft Systems (UAS)	Atmospheric science; greenhouse gases	To quantify and monitor greenhouse gas emission mitigation efforts in agricultural systems using Unmanned Aircraft Systems (UAS) (e.g., drones), with a current focus on commercial rice farming. To develop UAS capabilities for collecting earth science data, both with high resolution imagery analysis and lightweight gas sensors.	Challenge: Identifying sensors It is difficult to find sensors that are lightweight, small, inexpensive, and collect accurate data. Efforts to overcome challenge: A) They may borrow sensors from other laboratories, B) They will test less desirable sensors (e.g., may not be as accurate) that are affordable to determine if they are appropriate for their science needs, and B) Would be helpful if X-DOMES had a search option to identify sensors according to different requirements. Currently, it is possible to only download metadata about specific sensors. (i.e. user may want last week's earthquake data from Japan, not sensor 2345.)	Data discovery challenges, like CHORDS and CHILL, are about getting the right data from the requisite sensors. It is difficult to get data from the sensors Requests: Given that they are still in the design phase and scoping (based on funding) these may well change but hypothetically and ideally: Portals:A CHORDS real-time portal for monitoring SSEDD enabled science UAS in an XSEDE Jetstream virtual machine available for anyone to use.A UAS data portal providing access to collected source data (dependant on finding a DAC that is interested in prototyping UAS data hosting - none currently exists) DataRaw image files (RAW and JPG, RBG and Infrared bands)Stitched image files (JPG, GeoTiff, *.kml, Emotion project files, .LAS (point cloud), potentially a new open source 'stitched image project' file funding dependant)Raw text data logs of gas sensor voltage levels (likely in sensor manufacture proprietary formats)Text based logs of calibrated, error corrected, and unit scaled gas sensor readings.A multi-dimensional 'UAS data sensor output file' - To be determined/evaluated/developed as there currently is no standard way of encapsulating flight telemetry, with instantaneous sensor point data, with multi-channel imagery.Data plots and visualisations - web visualisations(html with D3), GeoTiff, *.kmlVideos (mp4) The interviewee requested: a search option to identify sensors according to different requirements. Currently, it is possible to only download metadata about specific sensors.	Data Format: geotiff, LAS, point cloud, *.csv, geojson, Volume (size): <5GB Velocity (e.g., real time): Data is collected from each fly over by the UAS	Ardupilot SSEDD Open options, also Emotion	Ideas for future but not known yet

Magnetosphere-Ionosphere-Atmosphere Coupling (MIAC) project_Gjerloev_out	Magnetosphere-Ionosphere-Atmosphere Coupling (MIAC) project	Atmospheric and Geospace	To provide complete global and continuous electrodynamic solution of the atmosphere. Solar wind (plasma flowing away from sun), magnetosphere, and upper atmosphere (e.g., ionosphere) interact in complex fashion to affect the Earth's space (e.g., space weather). To understand how these interactions work, it is important to be able to determine the complete electromagnetic solution of the auroral ionosphere.	Didn't have or have access to the required data sets previously. Now, with the existence of 3 data systems, they do. Remaining issue related to CI is maintenance of these systems.	Big Data - 1 TB per year, 20 GB of data per data.	They are combining data from the three different systems, so the data are all different. SuperDARN: collected data generated from 17 radars around the world; AMPERE: data are generated mostly from satellites; SuperMAG: data are sourced from more than 300 magnetometers located around the world (e.g., about 100 nations). All are vector measurements that represent time series data collected at different locations.	Developed at APL	For the data format - CDF, but there are none for the community
Mapping river migration from Landsat imagery_Schwenk_out	Mapping river migration dynamics from Landsat imagery	Geomorphology; global change; geographic location	To understand the migration of rivers through time using Landsat satellite imagery of rivers. Big questions ultimately are: How do humans impact river dynamics? (And how are humans impacted by migrating rivers?) How might climate changes and sediment dynamics affect river migration?	- processing the images to mark what is water and what is land is manual and very time consuming. He thinks he could develop an algorithm to do it. - No appropriate repo exists: he would like to share all of his data products, intermediate products, and tools, but doesn't know where/how. He had to develop all his code, and after he shared it, many people used it, and he was also told someone else had already developed that code (lack of sharing leads to redundant effort) - currently requires a supercomputer, but he would like to make data and processes desktop accessible - he thinks this is tractable. - having standards and defined methodologies would have helped him - data products are large enough that sharing them within his team, and externally, is hard. - no good visualization approach for showing the quantified data compactly.	He is able to access the data he wants from Google Earth Engine API and/or Earth Explorer (USGS product). This was not noted as a challenge.	- Data Formats: Images are downloaded as a *.tiff, he stores data as *.mat and *.m (matlab scripts). Intermediate data products that are raster images that may be in any format (e.g. *.tiff, *.jpeg, *.png, etc.). - Data Repositories: Google Earth Engine API, USGS Earth Explorere (Landsat archive) - Data Types: Landsat satellite imagery (input), data on river position and classification (width, cruvature, migration, etc.), maps of river - Data Volume: Downloading files from Landsat are about 250 GB, and he processes these down to make them smaller. Processed final map image and intermediate maps for the Ucayali River map are about 2 GB. He is unsure how much data volume would be required to make maps for all of South America. - Data Velocity (e.g., real time): Ranges from a few MBs to a few GBs a day. He averages about 0.5 GB a day. Not listed as challenge - Data Variety: not a challenge - Data Veracity: relies on USGS processing of Landsat.	Python and Matlab scripts that he wrote. Google Earth Engine	Lack of standards is a challenge
Merging and using seafloor observations_Rubin_out	Merging and using diverse datasets of seafloor observations	Volcanology, oceanography: general, geochemistry	To take observations from oceanic submersible vehicles and construct a geospatial time-series evolution of the sea floor. The originator's goal within this larger multidisciplinary project is to understand how volcanic landforms form over time, and how volcanic ecosystems work.	- Determining the precise location of submersible at a point in time for images/samples is difficult, error prone, has uncertainties that are not clear, and is often manual. But is critical for relating multidisciplinary datasets, or sharing data with other disciplines, and for building time series. A tool to help assign coordinates to data/video would be helpful (and improve based on expert opinion through time), as would computational experts that can help with sharing workflows and automating some components. - changing technology, e.g. needing to digitize VHS tapes. - searching 100s of hours of video and thousands of stills that have virtually no annotation. Finding video/images of interest is time consuming and manual. Would benefit from a more standard way to tag pictures and video for later review, as well as best practices to support consistency. (scientists take notes on observations from video in real time, but it is not clear what format these notes are in) - ArcGIS is the primary tool for his work, but uses a proprietary format that is hard to share or merge with colleagues that don't use ESRI.	- searching 100s of hours of video and thousands of stills that have virtually no annotation. Finding video/images of interest is time consuming and manual. Would benefit from a more standard way to tag pictures and video for later review, as well as best practices to support consistency. (scientists take notes on obsevations from video in real time, but it is not clear what format these notes are in)	- Data Type: video and still imagery of the seafloor - Data Source: collected from various submseribles, AUVs, ROVs, towed cameras, etc. There are also sensors on the platforms. - Data Formats: ArcGIS internal, NetCDF, CSV - Data Volume: 100TB per expedition, with 1-4 expeditions per year. 95% is video/imagery	ArcGIS, a bit of QGIS, Fledermaus (Google Earth and ArcScene are alternatives), VLC video imaging and editing, Excel.	- WOCE quality assessment convention and flags. - Observation notes are taken using a standardized language that is codified in the literature. Ditto for classification schemes. - GPS navigation and reporting datums
Metadata Database for Physical Samples_Hangsterfer_out	Metadata database for physical samples	oceanography; geochemistry	To have a database for curated, oceanographic samples that was both user-friendly and curator-friendly	- she currently has to upload sample metadata into 3 different registries. They do not use the same metadata format, and reformatting is manual and time consuming. - four catalogs exist: all have strengths/unique roles, no one is clearly best/sufficient on its own. - one of the catalogs is hard to edit - have to reload all data.	- none of the catalogs provides all the discovery features desired. Some desired elements: "look at" a large collection spatially, not too many hits when search on descriptors, see all samples from a cruise	- Data Repositories: NOAA IMGLS, UCSD Libraries Digital Collections; SESAR; the SIO Geological Collection web portal) - Data Formats: Flat files. *.xls. Associated files: images *. jpeg, analytical information about the samples. She would like to add *.avi for moving x-ray movies (e.g., CT technology – flying through the sediment core). - Data Volume (size): Up to 1GB per cruise with 130 cruises currently online. As more work is done on current samples and if new analytical equipment comes into the lab, this upper limit could increase. - Other characteristics not noted as challenges	Excel	The different repositories have different metadata templates

Moorea Coral Reef LTER_Edmunds_out	Moorea Coral Reef LTER	Community Ecology, Marine ecology, Ecosystem science; coral reefs	To document changes in coral communities over different time scales, including multi-decadal, and understand the causes; what is changing and why, and what is the vulnerability/resilience of coral reefs to different disturbances.	- remote field site with limited electricity & internet. Data must be moved off the island on physical storage media. - Need some kind of interface software that allows less accomplished computer/software people (like himself) sift through, conceptualize, and access the “fire hose” of data. There is a large chasm between the human capacity to think and make sense of the data, and the capability of the project to generate large quantities of data. - funding stability	- Too much data - hard to how to "sift through, conceptualize, and access" the huge amount (variety and volume)	- Data Repositories: BCO-DMO (NSF) and the Moorea data management program. Data used in manuscripts are identified with DOIs. - Data Sources: most data are field generated, but includes syntheses with other datasets from colleagues for context. Remote sensing, and in-situ field data. - Data Types: photos, counts/percent cover data from photos and transects, and measurements from biopsies of coral. On multiple spatial scales (from remotes sensing on thousands of km scale to small). - Data Formats: he uses .jpg, are upgrading to RAW proprietary photo format for higher resolution. *.xls, .mat. - Data Volume: 10s of GB per year from .jpg photos - Data Velocity, Veracity/Qualty: not noted as issue	- Analyze data using MATLAB, R, Systat, SPSS, and PRIMER.	- Data used in manuscripts get DOI. - They rely on broader scientific domain literature for definitions of benchmarks and best practices for determining the quality of their results and reporting
NARCCAP Data for City Management_McGinnis_out	NARCCAP data for city management	Global Change; Policy Sciences	To supply a city's managers (TX?) and others with future scenario temperatures based on the NARCCAP data, so that their team can make city infrastructure decisions.	- Large data transfers not easily supported, so end up sending hard disk drives to supply data - Need tools for extracting the relevant data (or subsampling) based on required info, such as location - Need format converters - Downloads are challenging - different internet browsers, security constraints - Analysts choose their own S/W. Different S/W support different data formats - Because they assumed the data would never change, it was not versioned and there were no plans to update it	- extracting the relevant data from the NA dataset - Analysts choose their own S/W. Different S/W support different data formats - different OSs have different and less or more tools to work with projected coordinate systems (it doesn't use lat/long - Data is in 5-year chunks - 30 day/month calendar vs. 365 day/year cause synching issues - Units differences, conversion needs	netCDF	User dependent, in this example case: ArcGIS	CF metadata specifications on NARCCARP site
Permafrost_Model_Validation_Schaefer_out	Permafrost model validation	permafrost; carbon cycling	To gather data on permafrost characteristics from many high latitutde locations to validate and improve a carbon cycle model.	- researcher must access and combine data from many studies. Desired data are in Arctic Data Explorer, but are discoverable with many different search terms, making searching time consuming and laborious - metadata is either insufficient or time consuming to assess whether data meet the researcher's requirements - data are in different formats, and need to be converted (to NetCDF)	Use case has extensive and detailed description of desired search and access workflow.	- Data Sources: time series datasets of permafrost temperature. - Data Repositories: Arctic Data Explorer (ADE) at National Snow and Ice Data Center (NSIDC) - Data Formats: netCDF (desired), Excel (often provided), other heterogeneous formats - Data Variability: format variability is a challenge - No issues noted for data volume, velocity, veracity.	BCube broker; - SiBCASA (combined Simple Biosphere/Carnegie-Ames-Stanford Approach) Terrestrial Carbon Cycle Model	Excel
Protein_Portal_Saito_out	Access of oceanic protein datasets	proteomics; ocean; microbial ecosystems; biogeochemistry	Create a community data portal that allows research scientists to discover where, when and in which organisms a protein/enzyme of interest occurs in the oceans through a bioinformatics analysis of large mass spectral libraries created from many oceanic samples.	- mass spectrometer data distributed among multiple repositories & different mass spec platforms will pose integration problems with different datatypes. - How to ingest up-to-date genomic information, because protein data relies upon genome data for production. - volume: .raw 10gb/day per instrument (~2-4 Tb/year/instrument), 1-2 instruments per contributing laboratory	Assuming proposed system is developed: - User comes with name of a protein or sequence to the portal - User enters parameters (ocean basin, year, etc) - User hits Go (submits form) - System matches name of the protein against a database of proteins System reports the protein found in the database along with information about the protein like where they were found, the number of spectra counts, geospatial & temporal reference frames (in table or visual [ODV form]), and the exact sequences - System reports displays results in a Map and a List (spreadsheet output) with quality indicators of the data and related metadata. - User evaluates results and can rerun the search with updated parameters - User can download the result set (eg Ocean Data View) Alternate flow: - User comes to portal with a geographic search (i.e., wants to see everything present) - If the System found a match, it would show a table (map would be too much data) of identified proteins - Otherwise, the System specifies no data found for that geographic region (region being pre-defined unit)	- Data Formats: .raw proprietary mass spectrometer-specific raw datafile, .mzXML common non-proprietary format for mass spectrometry raw data, .mzML new common non-proprietary format for mass spectrometry raw data, .mgf older format for raw data, .csv, .xml, .msf proprietary output file format, .fasta amino acid sequence database - Data Types: mass spectra, identified peptides after mapping to genomic datasets, spectral libraries to reduce search time, gene annotations information, relative and absolute abundance information - Volume: 0.6-50GB per file - Velocity: 2-4TB/instrument, 1-2 instruments per lab, multiple labs	ODV	none mentioned

Rapid Real-Time Atmospheric Hazards_Chandra_out	Rapid real-time response based on atmospheric hazards, multiple space-time scales	Atmospheric studies	To be able to observe and model atmospheric observations at real-time and multiple time scales, in order that stakeholders can respond to hazards more quickly and effectively.	- network speed vs. data collection velocity - not fast enough to capture all weather phenomenon - Not enough sensors in critical locations - Younger analysts, scientists would like to see data stored in the cloud, not just public data centers	(These are about data collection but could also apply to saved data) - Geometric mapping of different data types and all of the different inputs coming at once is really hard to do - Figuring out time scales to take measurements at and then matching them up - for each new variable in each location, timeframe - Weather effects can knock out ability to collect data, causing gaps in data - Synchronizing data	Input: Binary formats from radarsSigmat format, the company who makes the radartimevariable, etc. Output formats:NetCDF, HDF, Universal format	MATLAB IDL VCHILL NCAR tool	none mentioned
Raw sequence data conversion_Wood-Charlson_out	Convert raw sequence data into biologically relevant information	Biogeosciences, Informatics, Oceanography: biological and chemical	After data retrieval, to effectively translate raw sequence data into a format that provides genomic and phylogenetic context. In order to understand complex Earth Systems, most sequence-related data sets require a workflow to process raw data into a format that has meaning in the context of biological interactions (e.g. How microbes affect Earth systems, each other, and vise versa). The general idea is to find “who is where and what they are doing” related to sequence data.	- Need documentation standards - large data sets require nonlinear increases in computation - The community needs a common resource that can be used for most data scales and projects that can be used to standardize work between small and large labs.	•Database limitations – microbial community members may not be represented (may not exist) •Large portions of seq data sets are unknown/unidentifiable (have never been identified and are discovery-based) •On-going community data submissions are slowly expanding reference databases; it takes while to get things QC’d and is out of our hands •Major databases are working to clean up their data repositories—they are always improving	•.fasta common non-proprietary format for basic sequence data •.fastq common non-proprietary format for sequence data linked with quality data •.tsv •.csv •Outputs may vary depending on the tool and may have different file extensions that limit compatibility		Stds determined by Genomic Standards Consortium: MIMS/MIGS and MIxS
Reconstructing ancient rivers sedimentology in Big Horn Basin_Hajek_out	Reconstructing ancient rivers sedimentology from the Paleocene-Eocene thermal maximum in Big Horn Basin Wyoming.	Hydrogeology; Global change	The goals here are to use the stratigraphic record to infer past changes on the Earth’s surface related to continental rifting.	- Data are complex (field data, photos, samples, etc.), heriarchical, and have relationships between them - no system does this. Managing time is also challenging. Do not have any standards for how to handle. - Funding lag for long-term data storage. NSF - unsure about, NIH has promised to fund some. Want it to sustainable, beyond on NSF funding cycle.	- Data is stored both in SEAD, an NSF -funded project online database and also in SEN (an RCN). Also GeoPRISMs. But the PIs data is stored in lots of diverse locations. This seems to be on purpose because Kyle Straub is a test case (for what?) - No obvious way to combine experiments to numerical models (models are often in CSDMS) - Funding lag for long-term data storage	DEMs of topography in ASCII, xy attributes. Binary data stored by instrument, often proprietary format.	R MATLAB Excel Delft3d	Some SEASAR, IDEA Complicated story
Repeatable and Reproducible Geoscience Research_Malik_out	Repeatable and Reproducible Geoscience Research	Reproducible Research	reuse of data models and data preparation workflows, developed across several hydrology laboratories, and which must be reproduced for easy use by other members of the community.	Many hydrologica models have been developed, but the models and workflows using them are not easily reused. Geounit methodology is a solution needs adoption.	If tools don't support the mechanisms for annotation, it's difficult to get buy-in to use the annotation	Data Types and Repositories: -ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt -NCDC Global Historical Climatology Network (GHCND) database -Meteorological data from NCDC –GHCND -Topography (shape file/DEM) from USGSHyDRO1K http://www.usgs.gov/ Data Formats: Mostly text, time-series data. Example: netCDF, .csv, ,grib, Data Volume (size): Input Data: 1628.14 MB. Output Data: 1002.13	Graphviz (.gv)	netCDF BagIT
Science data system infrastructure_Law_out	Science data system infrastructure	Informatics	Provide the infrastructure, capabilities, and support for science discovery. Enhance science discovery by providing a data system that can support the science data life cycle. Allow for improved decision-making, and public education and communication.	- Project Management: budgeting and scheduling, finding the right talent, keeping up with emerging technologies, staying technical as much as possible. There is more funding than people to do the work. - Identifying end user requirements: scientists don't always know what they want. - Getting community buy-in - Many different data formats and metadata. People use different terms to describe relationships between data. This makes data systems architecture very complicated.		Data Formats: Many - depends on the client. This is a challenge. plus Geotiff for visualization	Lots, including for cloud computing	OGC Web services

Sediment transport rate field studies_Hsu_out	Finding data for sediment transport rate field studies and related resources	Hydrology; earth surface processes	•Better understanding of the relationship between stream flow and bedload sediment transport rates. •Flood hazard mitigiation and better understanding of longer term river evolution	- Ontologies are not built out effectively—including the vocabulary necessary for the queries - Processing queries requires concept expansion through an ontology. - Quite simply the sedimentary geosciences community is learning how to make these datasets available for wider use - Conversations between the data producers and the data publishers has started, but is not complete yet - In steps 4 and 5 of the basic workflow, the predicate 'is close to' is also complicated in detail, because resources that are related to the same catchment area or drainage basin as the sediment transport result from step 1 would be of interest at much greater distance than studies that were not so associated.	There are portals for very close data but not exactly what this use case is seeking. We're focused on bedload sediment. USGS is one producer of the data, and there are many others, but we need help aggregating these separate sources and users in order to make them effective over the longer term.	Data Sources/Repositories: •USGS databases (but this is mostly for suspended sediment as opposed to bedload sediment) •Disparate, small databases that are not linked because they are on personal servers or FTP sites Data Formats •CSV and text files are the norm •For the smaller databases, something like a text format •Volume (size) •Non-seismic data is pretty small; on the order of megabytes handful	OGC Geography Markup Language (GML)	OGC O&M
Seismology_CHORDS_Vernon_out	Streaming real-time data for seismic early warning and monitoring and environmental monitoring	Atmospheric processes; seismology; natural hazards	Develop a real-time streaming network for seismology and atmospheric sensors and cameras (data and processed products), for use by researchers, policy makers, first responders. Specifically, for Seismology and other areas: 1) Make it easier to bring in data in real-time. This enhances usefulness by getting the scientific analysis out of it faster and earlier, by making the best quality measurements. 2) Generating more, better data products that more accurately reflect the actual weather phenomenon. 3) Capture enough data to get big picture of environment with high-quality timing (correlated timing between sensors)	- Funding is insufficient to cover the number of sensors (don't have meteorological sensors at all monitoring sites) and the sophistication of algorithms he would like. - Need to figure out a way to build a sustainable network that's well engineered and migrated into a sustainable infrastructure. A long term maintenance model is needed.		- Data Sources: variety of in situ sensors giving images, meteorological, seismic, motion accelerometer, and state of health data. - Data Formats: miniC (a seismology format). There are MATLAB interfaces for this. Data formats are described on IRIS/DMC web site. - Repositories: IRIS DMC - Volume: 5TB/yr US, 10TB waveform for whole project. - Velocity: real-time (2 sec latency) - Variability: lots - Veracity: high	Most used: Antelope, MATLAB. Others: Fledermaus (3d 4d vis), ObsPy/Pyton	none
Strabo data system_Tikoff_out	Developing the Strabo data system	Structural geology; Geographic location; Tectonophysics	Develop a data system for field geological data (thin sections, maps, and empirical data) collected on multiple scales - a lack of data availability is limiting scientific progress.	- it is hard to acquire data needed to support research. This is inefficient. There isn't a culture of data sharing. - data are lost or unusable (not fully documented) when a researcher retires or dies. - ArcGIS is good but not sufficient for storing and analyzing their spatial data.	- not described, but it is implicit that they exist and are important.	- Data Source/Type: Thin sections, maps, and empirical data collected from a wide selection of scales used by field geologists. A geologic map at the kilometer scale, a detailed map at the 10's of meter scales, individual stations at the meter scale, thin sections at the millimeter scale, and then possibly even an electron miscopy sample at the sub millimeter scale. At any of those scales you could have data points, relationships, or even images linked to them. All of them are spatially referenced and are noted as such in the system - Data Repositories: data systems stored at Kansas. - Data Format: database will be Neo4j (graph database), inputs are ArcGIS or QGIS. Normal and annotated images as JPGs. - Data Volume: ~100MB/project, individuals probably use 10's of MB. Not highlighted as an issue.	- currently, ArcGIS, but it is not sufficient. - The Strabo system is in development and uses:Neo4j for graph DB, PHP and Python for Strabo itself.	None listed
Stromatolite distribution in space and time_Peters_out	Stromatolite distribution in space and time	Genetics; Paleontology; Mineralogy	To use GeoDeepDive to compile data and data synthesis on the frequency (e.g., peaks and declines) of stromatolite occurrence in space and time as a source of information about the evolution of earth's surface environmental conditions.	- before GeoDeepDive: data on stromatolite occurrence data were distributed in diverse scientific literature, and difficult and time consuming to access & extract. - Continuing challenges for GeoDeepDive: CPU is limiting, require coding skills to use that many students do not have, data quality problems, not able to find all data in a paper, etc. - how to credit data aggregations from GeoDeepDive is not fully resolved - can list the papers, but not all contribute equally. Need community best practices around this.	- finding data on stromatolite occurences in published scientific literature - as literature is added to GeoDeepDive, may want to repeat the search, but not clear if there is an efficient way to retrieve just the new hits.	- Data Formats: PDFs - Data Types: stromatolite distributions in space and time, from published literature - Data Repositories: published literature collections - Data Volume: could be up to 100s of GBs. Problematic volume for personal computers. - Data velocity, variety, and variability not listed as challenges	GeoDeepDive (use case was provided by GeoDeepDive team)	Open source, standard formats, annotation provided by software toolkit. Eg. Stanford NLP
Supraglacial Lake Depths From Satellite Imagery_Pope_out	Reproducibly estimating supraglacial lake depth from satellite imagery	Cryosphere	To use Landsat imagery to determine how much water is being stored in lakes on the Greenland ice sheet.	- It is a challenge to link together project artifacts and keep it all updated and synchronized - individual datasets, code. It is not automated but manual. new DOIs are constantly created and then manually propagated. - Preparing data for sharing.	Limited field data available, sharing data standards are missing (problem for others who want to use data that exists), scripts may not match data	- geotiff - MATLAB .mat - .csv - NASA and USGS satelite data (Landsat)	MATLAB (mostly) GDAL	None

Top ten storms in the NARCCAP database_McGinnis_out	Top ten storms in the NARCCAP database	Global Change	Identify intense storms in the NARCCAP dataset occurring near Fort Collins, CO	If the researcher doesn't have access to the Supercomputer at NCAR, they are unable to process the big datasets, as they don't fit on the researcher's computer.	See CI challenges - need access to and time on supercomputers, plus skills to process large data	net CDF	NCL WRF	CF metadata specifications on NARCCARP site
Use of high-resolution commercial imagery by NSF investigators_Morin_use	Storage and use of high-resolution commercial limagery by NSF investigators	Topography; geomorphology	To complete high resolution (submeter) elevation maps of the polar regions that are useful for the science community	- large volumes of data must be stored, and hard to find funding for it. He would like to see a community cloud storage facility. - data for some regions must be requested and can take a day to several months to be delivered, and it not available to all (only NSF-funded projects and govt)	Discovery not mentioned as a challenge per se. More about access after discovery.	- Data Sources: satellite images, Digital Elevation Model output - Data Types: satellite imagery, digital elevation models output, imagery mosaics. - Data Repositories: National Geospatial Intelligence Agency (NGA) - Data Formats: *.ntf (image data used by defense and intelligence for remote sensing). GeoTIF for elevation data and some imagery products. - Data Volume: Petabyte a year. They hold about 3-5 petabytes. - Data Velocity: 4 satellites are collecting imagery, 24 hours a day - Data Veracity/Qualitu: is managed by NGA repository.	Standard geospatial software GIS and remote sensing). ARCGIS, Imagine, ENVI, GDAL and many open source packages.	
Volcano tectonic interactions on a global scale_Stamps_out	Investigating volcano tectonic interactions on a global scale	Tectonophysics; volcanology; natural hazards	To better understand the role active volcanism have on global tectonic movements.	- We need millimeter precision positioning data that spans the active volcano and area around the volcano, ideally in a cross-section. - Being be able to compile and assess a wide range of data in one digital location is challenging. - Logistical challenges to collecting new data: problem with data producers (where should instruments be placed? How will they be protected?) - The issue of discrete data versus continuous data—some sensors record continuous data, but others are inherently discrete. For example InSAR and geochemical observations are discrete. - Once all that data gets compiled there would need to be some kind of statistical analysis, and that is another way EC can help.	Desired workflow: integrated data access. (GeoMapApp is an example tool that allows you to pick a region and find data that's available. However GeoMapApp doesn't crawl web.) Old workflow - without integrated discovery:To go to each database Extract position for each individual volcano Pull information about geochemistry and see where the data is located Manually process the data It may require 5 years to assess data related to active volcanoes and tectonic deformation at an appropriate scale.	- Data Types: InSAR; Positioning sensors (like GPS); Petrology (geochemistry data); Temperature data; Volatiles from the off-gassing of volcanos; Gauge data; seismic data - Data Repositories: (further details on each given in the use case) United States Geological Survey Volcano Hazards Program.The Volcano Disaster Assistance Program (VDAP). Global Volcano Model. WOVOdat. Integrated Earth Data Applications. VHub. International Association of Volcanology and Chemistry of the Earth's Interior (IAVCEI). National Oceanographic and Atmospheric Administration (NOAA) Volcanic Ash Advisory Centers (VAACs) - Data Formats: GPS data in NMEA and BIMEX, Dataing data in ASCII, Geochemistry in ASCII and Excel, Temperature and GPS in ASCII, Seismic in SEED. - Data Velocity: some is real time (GPS and seismometers) otheres delayed mode. - Data Variety: see data sources. Also field collection vs real time sensor. - Data Veracity/Quality: GPS data needs processing, and QC is problematic. She only trusts published data.	GAMIT-GLOBK. GMT, generic mapping tools. GeoMapApp for visualization. Google Earth—there are a number of software that allow you to output into the KMZ or KML formats, which then allow you to visualize. VIZIT visualization software. Adobe Illustrator. The IRIS web site also has some good tools to view certain mechanisms. MATLAB, QCSH, Bok. GOT for statistical analysis. LaTeX for word processing. TEQC physical analysis program.	There are some types of standards, specifically for waveform data, as well as for GPS data, but there aren't specific standards for this specific use case.
Watershed_model_Chapela_out	A dynamical watershed model for concentration-age-discharge in a highly weathered tropical site (Luquillo CZO)	Computational hydrology (geophysics)	Improve knowledge of the relationship between chemical weathering, nutrient cycling and hydrology by developing model repositories and online access to data to support the hydrological parameters	1. Not all data are discoverable and accessible - LTER Catalog helps, but is not complete, so individual scientists need to be asked for their data. 2. The data require QC and manipulation that is time consuming. Some of the QC could be automated if community scripts were available. Variations in parameter names and units also made time consuming, and metadata was not always complete 3. There were gaps in the data, making it hard to find sites and timespans with complete data.	I also spent a lot of time finding the right parameters, as different terms are used exchangeably or change through historical records, or units that are common ground for hydrologists/climatologist, but were unknown to me. While usually I could solve that by looking at the explanation in the LCZO metadata, the LTER one is less clear, and sometimes I had to contact the data manager directly or look up several books. This could be easily solved having a hydro-Glossary or similar, as well as common units and abbreviations. Most of the problems I encountered can be grouped as: a) Missing values b) Values at sampling intervals longer than required by the model c) Values at sampling intervals shorter than required by the model d) Different units Part of this search is that she wants to find sites and timespans with complete data. Not just do data exist in that place/time, but is it complete.	- Data Sources: LTER Data Catalog, LCZO Data Catalog, published papers, *and* privately held, unpublished data. - Data Format: Mostly .csv files, few .txt and excel. The only data heterogeneity problem was with dates and with units. CDF (Computable Document Framework) format or documenting models that provides the basic theory, equations and numerical solution along with visualization of results - Repositories: Luquillo LTER data catalog, LCZO repository - Variety (multiple datasets, mashup): Data heterogeneity was an important challenge, both managing the variety and figuring out which are best suited to the model - Variability: Missing data (gaps) was a large challenge - Veracity/Data Quality (accuracy, precision): Reliability of some of the data was an issue, and could not have been spotted without hydrology expertise; asking the LCZO data manager and PI was needed to finally discard some data. - Data Types: All were time series, except for well logs and map of the site	- GitHub: a tool for sharing software and tracking versions. - GeoSoft model library - The model Maria based her customized model on, residing in GitHub - Maria's customized model, a changed version of the above. - Matlab and Mathematica: the environment in which Maria edited the model. - Skype: for communication	EML (Ecological Metadata Language) ISO 19115 metadata DOIs