DATA SCIENCE JOURNAL

Stocks, KI, et al. 2019. Geoscientists' Perspectives on Cyberinfrastructure Needs: A Collection of User Scenarios. *Data Science Journal,* 18: 21, pp. 1–15. DOI: https://doi.org/10.5334/dsj-2019-021

RESEARCH PAPER

Geoscientists' Perspectives on Cyberinfrastructure Needs: A Collection of User Scenarios

Karen I. Stocks[1], Sam Schramski[2], Arika Virapongse[3,4] and Lisa Kempler[5]

[1] Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, US

[2] Center of the Analysis of Social-Ecological Landscapes, Indiana University, Bloomington, IN, US

[3] The Ronin Institute, Montclair, NJ, US

[4] Middle Path EcoSolutions, Boulder, CO, US

[5] MathWorks, Natick, MA, US

Corresponding author: Karen I. Stocks (kstocks@ucsd.edu)

Cyberinfrastructure (CI) is a standard tool in the geosciences, but the creation of successful CI remains difficult, and expensive projects can have significant consequences for scientific communities if they do not result in success. In this paper, we present an effort to solicit feedback on cyberinfrastructure needs from a broad community of geoscientists by means of user scenarios to inform the National Science Foundation's (NSF) EarthCube program. The method for the user scenarios was semi-structured interviews, a total of 50 of which were collected from a broad range of scientists and analyzed. A wide variety of challenges were identified, with the most commonly articulated challenges being an inability to find data of interest in an online repository, the heterogeneity of data and metadata, the lack of needed software (which in turn drove redundant development of needed software in multiple groups), and insufficient or unstable funding for long-term cyberinfrastructure. While the user scenarios do not provide formal requirements in the software engineering sense, they do provide expressions of user challenges that, in many cases, are sufficiently detailed to inform high-level requirement development.

Keywords: Cyberinfrastructure; user scenario; use case; geosciences; EarthCube

1. Introduction

Geoscientists, and scientists in general, use cyberinfrastructure (CI) as a standard tool. Whether called CI, e-science, or information infrastructure, CI in the broad sense supports a wide variety of geosciences research and education. From small-scale local databases and desktop analysis tools to high-performance computing and massive online data collections and models, it is a part of many geoscientists' workflows. Currently, there are substantial efforts to develop and maintain CI for the geosciences such as EarthCube in the US (EarthCube 2018), ENVRI Plus in the EU (Asmi et al. 2017), and the Australian Ocean Data Portal (Open Access to Ocean Data 2018).

Unfortunately, creating successful CI remains challenging and far from straightforward (e.g. Dooley et al. 2006; Finholt & Birnholtz 2006, Ribes & Lee 2010). Cyberinfrastructure projects can be expensive, and the cost of failure is more than monetary: a major failure can impact an entire research community's perspective of the value of CI development for the future. Because the potential exists for very significant gains or losses, it is important for programs and funders to assess the needs of their intended user community accurately. Building a desktop data management tool for a community that uses cell phones or tablets to collect data in the field will prevent success, even if the desktop software is designed and performs wells.

There are multiple avenues for capturing end-user needs, including informal dialog with the research community, training expert advisors, and the application of questionnaires. Here, we present an effort to solicit feedback from a broad community of geoscientists, using semi-structured interviews to collect a set of user scenarios to inform the EarthCube program (EarthCube 2018). EarthCube was initiated by the National Science Foundation (NSF) in 2011 to develop cyberinfrastructure to support geoscience research and improve access, sharing, visualization, and analysis of geosciences data and related resources. EarthCube's goal is to enable geoscientists to tackle the challenges of understanding and predicting complex and evolving

solid Earth, hydrosphere, atmosphere, and space systems. It is a community-governed effort, with priorities and directions informed by working groups, community-contributed white papers, charrettes, end-user workshops, booths and town halls at geosciences meetings, and broadly distributed questionnaires, as well as user scenarios.

The user scenario work was conducted by the Use Case Working Group (UCWG) within the EarthCube Technical and Architecture Committee and serves to supplement end-user outreach by collecting detailed descriptions of geoscientists' workflows and how they interact with CI. (The term 'use case' was originally applied to these products, because the EarthCube community had a history of using this term, though they are more similar to Carroll's (2000) 'user scenarios').

The purpose of the user scenarios collected by UCWG was to understand the CI needs and priorities of the geosciences community. Geoscientists were asked to describe their research, and identify those parts that were difficult or impossible in a way that new or better CI could help. These user scenarios are structured, thorough, in-depth expressions of how each scientist does his or her work, and what the pain points are CI might be able to address. From the user scenarios we summarized the common needs and concerns expressed across different science domains. In addition, we reflect on the utility of semi-structured, interview-based user scenarios as a mechanism to assess the CI needs and priorities of a scientific community.

2. Methods

2.1. User scenario template

The UCWG collected user scenarios from geoscientists using semi-structured interviews. User scenario collection is a well-established practice in the design of software and computer systems (Carroll 2000). A user scenario is a detailed description of what users do with a software or CI component and, more importantly, why they do it. It is longer and more informative than a user story, which is a short, simple description of a feature told from the perspective of a person who desires new capabilities within a system (Cohn 2009). User scenarios outline the motivation, main goals, and workflows of users.

An ideal user scenario provides an explanation that defines the context. It helps to describe: What does the user want to accomplish with our product? How is the user going to achieve their goals? Why does the user choose this product over other available options? Oftentimes, user stories provide detail on both the users' processes and their characteristics (Assistant Secretary for Public Affairs 2013). In our case, we did not focus on a particular product, but on users' current use and needs for new cyberinfrastructure overall. A key focus was understanding the work and challenges of domain experts in the field of geoscience from their perspective. Essentially, these user scenarios are in-depth expressions of how each scientist does or would like to do their work, which parts are difficult or impossible, and, finally, what must be improved by CI in order for that work to be possible or easier. To attain the desired specificity, each interviewed scientist was asked to describe and detail a specific project or research direction in which they were personally involved or interested.

A single, structured format for documenting user scenarios was used to promote the efficiency and consistency of data collection, so that the resulting information could be more easily compared, searched, and analyzed. The UCWG developed a user scenario template based on the Basic Use Case Template from the Rensselaer Polytechnic Institute's Tetherless World Constellation (2014), expanded with technical detail from the NIST Big Data Working Group (2017) Use Case template V1.0 and adjusted for EarthCube needs with review and input from the EarthCube Science Committee and Technical and Architecture Committee. The template was designed to collect information about a single scenario of choice by the respondent (i.e., a specific project or research direction). The template included questions about the scenario's overall scientific motivation, goals, and importance; the people and systems involved (including existing CI); a detailed workflow; the primary challenges; the desired outcomes; technical details about the data, software, and standards used or desired; and supporting documentation (**Table 1**).

2.2. Selection of interviewees

The study sample consisted of self-identified geoscientists, who also represent the target population for the EarthCube initiative. All but one individual were from US institutions. Individuals were selected opportunistically by recruitment through fliers distributed at the 2015 American Geophysical Union Fall Meeting meeting, announcements on the EarthCube website and at EarthCube meetings, direct requests from the UCWG members to researchers, requests for interviewees to recommend other potential participants, solicitations of past EarthCube end-user workshop chairs, and purposive sampling of scientists who were principal investigators or key contributors to active geoscience research projects (mostly Earthcube

Table 1: User Scenario Template. See Stocks et al. (2019) for an annotated user scenario template with additional usage tips and example responses.

Summary Information Section

Use Case Name

Contacts: Roles for the contacts are taken from the ISO 19115 CI_RoleCode vocabulary, and at a minimum should capture the interviewer (as Author) and interviewed scientist (as pointOfContact/originator).

Link to Primary Documentation: a single reference that describes the scenario, if it exists. Additional related references are captured below.

Permission to make public? (Yes/No); **Permission granted by; Date permission granted**

Science Objectives and Outcomes: overview of the scientific goals and importance of the scenario.

Overarching Science Driver: high-level scientific impetus. Can referencing agency priorities or strategic science plans documentation as relevant.

Scenario Detail Section

Actors: key people and/or systems involved in the project.

Preconditions: preconditions, requirements, assumptions, and state changes necessary for the scenario to be executed.

Critical Existing Cyberinfrastructure: existing data repositories, software, etc. needed.

Measures of Success: the important outcome or product if the scenario workflow is completed.

Basic Flow: steps to be followed in doing the user scenario. Often referred to as the primary scenario or course of events.

Alternate Flow: any alternate workflows that might occur, e.g. to handle error conditions.

Activity Diagram: a picture or flow chart that captures the major workflow steps, actors at each step, inputs and outputs at each step, and optional alternate paths.

Major Outcome and Post Conditions: conditions that will be true of the state of the system after the scenario has been completed. Including what happens with data and other products after the project finishes.

Problems/Challenges: any significant or disruptive problems or challenges that prevent or interfere with the successful completion of the activity. For each one, list the challenge and who/how it impacts; what, if any, efforts have been undertaken to fix these problem; recommendations for tackling this problem; how the larger community can address this problem.

References: links to other relevant information such as background, clarifying and otherwise useful source material. Include web site links, project names, overall charters, additional points of contact, etc. This section is distinct from Primary Documentation, which just describes the particular scenario.

Notes: Any additional important information.

Technical Section

Data Characteristics: describe all the data involved in the scenario, both existing and desired, as follows:

Data Source(s)

Data Format(s)

Volume (size)

Velocity (e.g. 2TB/day)

Variety (e.g. sensor data, model output)

Variability (e.g. differences in site density across studies)

Veracity/Data Quality (accuracy, precision)

Data Types (e.g. sequence data, core images)

Standards: any standards that were followed for the cyberinfrastructure resources, even if already mentioned above. Standards can apply to data, models, metadata, etc.

Data Visualization and Analytics: analysis and visualization capabilities needed for the scenario, whether existing or desired.

Software: For any important software used, describe the important characteristics (source, language, input format, output format, CPU requirements, etc.).

Metadata: Provide a link to, or include, any relevant metadata adding additional detail and context to the dataset(s) described above.

projects). We sought to foster a reasonable coverage across geosciences domains and did not interview more than one referral from an interviewed scientist. Interviewees were asked to provide a user scenario that met one of the following criteria: 1) an earth science research project that had been completed or was in process; or 2) an idea for a project that had been elaborated upon extensively, even though it was not yet active.

2.3. Interviews and data collection

Initially, scientists were asked to self-report using a user scenario template that was designed as a structured questionnaire—a tool that presents a set of identically worded questions in the same order to each participant. However, participation was low and completed scenarios lacked the desired level of detail. As a result, the authors changed to a semi-structured interview approach. Semi-structured interviews are based on an interview guide, so that the same list of questions and topics are covered. However, an interviewer can change the wording and order of the questions, and add follow-up questions to deepen responses. Semi-structured interviews allow for a high level of detail and completeness of coverage (Bernard 2012: 182–183). In this case, the user scenario template provided the guide for the semi-structured interviews.

Interviews were conducted primarily by authors Virapongse and Schramski, who have formal training and experience with interview approaches. (The self-reported scenarios that had been collected were either repeated as semi-structured interviews or not included in the results.) Interviews were conducted primarily via phone or video chat, with one interview conducted in-person. Most interviews lasted about 1.5 hours. At the beginning of each interview, the participant was given an introduction to EarthCube and the user scenario effort. We presented an informed consent (Stocks et al. 2019) to all interviewees to describe the research structure and goals, and ensure that they understood their scenarios would be made publicly available and that their participation was entirely voluntary. The informed consent form was sent in advance of the interview, and consent was requested during the interview and recorded in the user scenario document. All interviewees gave their consent to participate.

Prompting and verification methods were used to collect as much detail as possible, ensuring that user scenarios focused on a single specific research scenario (vs. hypothesizing about the needs of the field in general), and verifying the accuracy of their responses. For example, interviewers used different questions to elicit responses about the same topics, administered follow-up questions, and reminded the interviewee about the intended scope of the interview.

During the interview, the interviewer collected notes which were further organized and edited for clarity by the interviewer after the interview was complete. The interviewee then received their completed user scenario for revision, review, and verification. User scenarios averaged approximately 1500 words long, but varied substantially in length. Finally, scientists were asked to characterize the scientific domain of their scenario based on the set of index terms used by American Geophysical Union (2018) to categorize abstract submissions to its annual meetings and manuscript submissions to its affiliated journals, although we allowed for the option of adding free keywords. The AGU term list is an imperfect vocabulary, in particular for oceanographic and biological sciences, but it is accessible and easily understood. The authors filled in domain science keywords for user scenarios when the interviewee did not provide them.

2.4. Data analysis

The completed scenarios were summarized based on several aspects. First, the authors extracted a list of CI needs or challenges for each user scenario. Second, the specific data formats, software, and standards (in the broadest sense) mentioned were listed and enumerated across the user scenarios. This information, along with summary information about the scientific focus of each scenario, was captured in a summary matrix (Stocks et al. 2019).

Next, we conducted qualitative analysis by grouping similar challenge statements together, iteratively revising a general theme that described each group of statements, and finally, identifying the main (i.e., most frequently identified) CI challenges expressed by interviewees. User scenarios often had multiple challenges, and groupings were not necessarily mutually exclusive.

3. Results

A total of 52 user scenarios were collected, and 50 were included in the analysis. One interview was excluded because it was incomplete (the interviewee did not complete the final review and approval), and another was excluded because it focused on a hypothetical research scenario instead of research the interviewee had actually conducted or wished to conduct. Approximately 60% of the interviewees provided keywords for the specific scientific domain(s) of their scenario.

There was a wide breadth of user scenarios. For example, they included: an attempt to develop and build four-dimensional geologic models of use to researchers in diverse geosciences communities; management of data developed from CO_2 measurements in seawater; and understanding how, where, and when magma accumulates underneath active volcanoes, and what is involved in the magma eruption process.

The sample group of 50 individuals worked predominantly at US institutions (98%), with the majority (76%) from academic institutions, 22% from government, and one use case (2%) from a not-for-profit research organization. The sample group was composed of 62% men and 38% women. The use case matrix shows the domain keywords used to describe each scenario. Overall, geology is the most strongly represented, but there are several user scenarios each for oceanography, hydrology, and the atmospheric sciences. Many were multidisciplinary or otherwise difficult to categorize.

3.1. Cyberinfrastructure challenges

3.1.1. Data challenges

By far the most common CI-related challenges expressed in the user scenarios were related to finding, accessing, using, reusing, and sharing data (**Table 2**). **Figure 1** shows a word cloud made from the extracted statements of CI challenges in which the size of each word is approximately proportional to the number of

Table 2: Summary of cyberinfrastructure challenges expressed in the 49 use cases. The percents do not add up to the category totals because 1) one use case can express challenges in more than one subcategory; and 2) challenges expressed by three or fewer use cases were not listed, but were included in the category counts. Percentages are absolute not relative.

Data Challenges	
76%	Data Access/availability
28%	Data not online
18%	Data in multiple online sources
14%	Hard to search for desired data in online source
12%	Important relationships between data in multiple sources missing
8%	Hard to find/access data in publications
8%	Sharing data is difficult/lacks incentives
32%	Data variety, diversity, and heterogeneity issues
24%	Data format diversity
8%	Semantic variability
12%	Integrating different data types (discrete vs continuous, sensor vs 4D model, etc.)
Other	
18%	Total data volume
16%	Needed data does not exist (e.g. not enough sensors, or gaps in the data)
14%	Insufficient or uncertain data quality
14%	Insufficient metadata
Non-Data Challenges	
36%	Software
30%	Desired software does not exist
12%	Desired software exists, but is not accessible/reusable
28%	Best practices, protocols, standards, other guidance needed
26%	Funding challenges, especially long-term sustainability for CI
16%	Networking, Storage, CPU
12%	Access to informatics/computer science expertise

Figure 1: A word cloud created from the summary of cyberinfrastructure challenges extracted from each user scenario. The size of each word is approximately proportional to the number of uses of that word. Common words and numbers are not included.

times it appears in a set of challenge statements. Though the exact phrasing of the CI challenge is subjective, making this image more approximate than exact, the importance of 'data' is evident.

The most common data challenges cited in 39 scenarios (78%) were related to data access or availability. More specifically, 28% of scenarios referred to data being unavailable in online repositories, either because no appropriate repository exists for the specific kind of data of interest, or because scientists generating the data were not using repositories. Another common complaint (18%) was that data were divided across multiple/too many online sources and were, therefore, time-consuming to find. Additional concerns around data access were that it was difficult or impossible to search data sources in the way(s) desired (14%), the connections or relationships between data in multiple locations were not expressed, so finding related data was difficult (12%), and it was difficult to find or extract data from publications (8%). A total of 8% of the user scenarios expressed challenges from the perspective of the data provider, noting that data sharing was difficult and proper incentives do not exist for data creators to make their data available.

The next most common data-related challenge was around data variety, diversity, and heterogeneity (32% of user scenarios). More specifically, 24% of scenarios reported challenges with having to work with or convert a diversity of formats or, for a few, working with proprietary formats. Conversely, as one user explained concerning a scenario, '[T]he 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).' Some user scenarios (8%) noted that semantic variability made data discovery challenging: different terms were used to describe data across datasets and portals.

A second common category of data diversity challenges was related to the complexity of the records. A total of 12% of scenarios identified integrating data from diverse sources as difficult. Ranging from in situ point data to images to 4D tectonic models and laboratory analyses, even within the geosciences these forms are often incommutable. Then there are the fundamental characteristics of the data, such as their expression as either discrete or continuous.

Additional data usability problems related to insufficient metadata (14%) and unknown or insufficient data quality (14%) (there was overlap in these two categories). Total data volume created difficulties in 18% of scenarios. The most common data size threshold was when datasets became too large to hold conveniently on a laptop computer (roughly 100GB+ per dataset). Finally, for 16% of respondents, the data they desired did not exist. Examples included a lack of sufficient sensors and significant spatial or temporal gaps in the data.

3.1.2. Software and tool challenges

Though data-related issues were the most common concern for the geoscientists interviewed, 36% of the scenarios highlighted a challenge related to software, which broadly included analysis tools, scripts, models,

quality assessment routines, and other code. In contrast to data, the primary concern was that desired software did not exist (30%). The kinds of software desired were diverse, and most were highly domain specific (e.g., a tool to add location coordinates and uncertainty to video from submersibles).

Access to existing software was a secondary but substantial challenge (12% of scenarios). Inefficiency was a concern here: 10% of scenarios noted that the lack of shared software was a problem because it created duplicate efforts, as code to carry out common tasks had to be reinvented locally.

3.2. Hardware and networking challenges

A relatively small proportion (16%) of scenarios cited a challenge related to bandwidth/networking (6%), access to high-performance computing (6%), or the need for more storage (4%). Though given that 18% of scenarios noted a challenge around data volumes, respondents may have underestimated their hardware and networking challenges; by definition there exists either a hardware (storage hardware not sufficient) or a networking (inability to migrate data) limitation when volume presents a difficulty.

3.3. Broader challenges

Interviewees also cited challenges that were not specific to technical cyberinfrastructure resource (e.g. data repositories and tools). These included the need to develop or adopt best practices, protocols, standards or other guidance (28%) and funding challenges (26%). Within funding challenges, the lack of longer-term funding sustainability and reliability needed for supporting community cyberinfrastructure were mentioned. Limited access to informatics or computer science expertise was also stated by 12% of scenarios as a challenge.

3.4. Existing practices: data formats, standards, software

The interview template included specific questions about data formats, software, and standards. In contrast to the previous section on the software-related challenges scientists faced, this section simply asked scientists what they were currently employing.

3.4.1. Data standards and data formats

The standards mentioned and their frequency of mentions from the user scenarios are listed in **Table 3**. The standards listed are from the point of view of the interviewees (i.e., we did not remove responses that are not technically considered standards). In the narrative parts of the scenarios, it could be ambiguous whether the scientist considered something to be a standard, so our counts may be imperfect. Of the 25 standards mentioned, six occur in more than one scenario, with Open Geospatial Consortium (an organization endorsing a family of standards) being the most common. The remainder were mentioned in only a single scenario. Many scenarios (26%) didn't specify any standards or specific data formats.

Table 4 shows the data formats mentioned. Just as with standards, what the interviewee considered a format was open to some interpretation and it was difficult to tally. Of the 25 formats mentioned, ten were mentioned in more than one scenario; the most common format, CSV, was used in 14 (28%) of scenarios.

3.4.2 Software

Table 5 lists software with more than one mention. Overall, there were 155 mentions of software, for an average of about three per user scenario. In these 155 mentions, 92 different kinds of software were named. While the overall number was quite large and contained numerous single mentions, there were a handful of commercial packages (MATLAB, Excel, ArcGIS) with wider use (7 to 17 mentions each). The second most common software was not a specific category but the catch-all category of custom code created in-house.

4. Discussion

The importance of user scenarios can be thought of in two ways. First, there is value in the qualitative features of user scenarios (Yin 2017), placing the CI use or needs in the context of a well-described research flow. Second, the individual quantitative elements or counts of features found within the scenarios are informative. The qualitative value in interviews has long been recognized by social scientists as robust, perhaps encapsulated most succinctly as the need for 'thick description' in social research by Geertz (1973), and inherent to ethnographies and unstructured interviews. Software development has similarly recognized the value of counts or tallying of content within a text, such as in the collected scenarios found in this study, and is often carried out using computer-aided text analyses (Short et al. 2010). This paper argues that the collection of detailed scenarios is a primary product of this work, not merely an intermediate step toward tallying and totaling. The latter are hallmarks of interview analyses focused on statistical outputs

Table 3: Standards mentioned in user scenarios. Note that this represents the scientists' reporting on the standards they use, and items like 'GPS' are included even though a technologist would not consider them a standard.

Standard	# of Use Cases
OGC	5
DOI	2
EML	2
GPS	2
iGSN	2
NetCDF	2
BagIt	1
CDF	1
CF metadata	1
CUAHSI WFS	1
DCAT	1
Excel	1
GCIS	1
iPLANT	1
IRIS	1
ISO19115	1
Memex	1
MIMS/MIGS	1
NOAA	1
Nutch	1
SEAD	1
SEASAS	1
SensorML	1
USDA	1
UTF-8	1
VIVO	1
WOCE	1

procured after numbers and digits have been spun from non-uniform textual content. This is not to say this paper eschews the need for identifying patterns; the results of this work show that many emerge across the scenarios.

4.1. Study sample

Most of the scientists interviewed were or had been involved in EarthCube in some way (e.g., had participated in an EarthCube workshop, meeting, or funded project). Efforts were made to recruit more widely, for example through flyers handed out at the EarthCube booth at AGU, but these efforts did not attract large numbers of volunteers. The bias this may have caused is unclear, but the user scenarios are likely to have oversampled geoscientists who are more interested and engaged in using cyberinfrastructure as part of their work than the average geoscientist, so they are perhaps more sophisticated CI users.

Advertising materials (fliers, EarthCube newsletter announcements, emails, etc.) all requested that practicing geoscientists talk about their own research. However, the respondents included several

Table 4: Data formats mentioned in user scenarios.

Format	# of Use Cases
CSV	14
NetCDF	11
MATLAB .mat	6
Excel	6
txt	5
ArcGIS/ESRI shapefiles	4
jpeg	3
tiff	3
tsv	2
SEED	2
xls	1
mzML	1
mzXML	1
geojson	1
geotiff	1
GIS	1
grib	1
HDF	1
HTML	1
IRIS	1
JSON	1
miniC	1
MSAccess	1
Pivotpilot	1
png	1
UTF-8 unicode	1

scientists involved in developing cyberinfrastructure and we did not eliminate these respondents. Given that the line between a technologist and a scientist is not always clear, and scientists are often in the position of building their own tools, these individuals have valuable perspectives. In general, these technologists were often more challenging than domain scientists who are not involved in CI development to interview, as they tend to generalize the problems in the field rather than share their personal experiences.

4.2. Challenges

The ability to find, access, and use data was the leading cyberinfrastructure-related challenge that the interviewed geoscientists expressed (76%). This suggests that geoscientists would be served by expanding and improving data facilities.

These results are consistent with the results of a series of surveys of potential EarthCube stakeholders (both geoscientists and CI developers) conducted in 2012–2014 (Cutcher-Gershenfeld et al. 2016). These surveys did not differentiate between data and software challenges, but asked overall how important and easy it is to find, access, or integrate multiple datasets, models, and software. Overall, the importance levels were high (the average was 0.76–0.87 on a scale of 0–1, with one being highest), but the ease levels were

Table 5: Software mentioned in user scenarios. Only those found in two or more use cases are listed. Note that the category 'in-house' is not a single software, but includes any mention of unnamed software/code developed by the group of the interviewee.

Software	# of Use Cases
MATLAB	17
In-house code	10
Excel	9
ArcGIS	7
R	5
Adobe Illustrator	4
Python	3
Google Earth Engine	3
IRIS/DMC tools	3
IDL	2
NCAR Tool, NCL	2
Fledermaus	2
Mathematica	2
ODV	2
Paraview	2
GDAL	2
VLC	2
Petrel	2
SQL	2
STRABO	2

markedly lower (0.30–0.41), indicating a desire for access to CI resources but an expressed belief that there were challenges in using them.

For geoscientists who stated that their data of interest is not online, it is possible that they are correct or that the data do exist online but they are unaware of where to find it. This is likely to be an infrastructure problem, but may also be an education or usability issue. It might be informative to dig deeper into this challenge to evaluate this claim: for the specific data desired by scientists, does a search of existing data resources and consultation with experts indicate that there are data resources suiting their needs? If a repository does exist, it is important to understand why a scientist may have ruled it out, perhaps considering it not useful or inaccessible in some way.

Furthermore, data discovery and access is generally an early step in the scientific workflow. It may well be that once data of interest are accessible and usable, the scientists will then have further challenges regarding the tools to work with the data, storage volumes, computational power, etc. As with data availability, scientists who state that the software they wish to have does not exist may either be correct, or unaware of existing code. If new software is needed, does it require further research and development, or is it simply a matter of effort or initial funding to share it? Relatively few user scenarios reported hardware or networking challenges. Data storage was a commonly cited problem, with the most common data size threshold being when datasets became too large to store on a laptop computer (roughly 100GB+ per dataset), perhaps suggesting the importance of cloud resources and co-located data storage and analysis tools.

A perennial problem for scientists in this study, as is often the case when discussing the development of cyberinfrastructure (Kee & Browning 2010), is the nature of funding. This dilemma was cited in more than a quarter of user scenarios. Given the inherent social and aggregative aspects of cyberinfrastructure and the dearth of low-cost maintenance options for these scientists' needs, this concern is likely to persist even as solutions are developed for the benefit of many.

4.3. Existing practices: data formats, standards, software

The responses regarding standards, formats, and software indicate a long tail of heterogeneous practices: for each category, the majority of instances were found in a single user scenario. Our methods may have undercounted formats and standards used, as we used an open-ended question to elicit responses. A more focused inventory-style questionnaire format, asking questions such as 'do you use R?' or 'do you have a problem visualizing data' may have prompted additional feedback.

Concerning CI planning, the results indicate that an architecture supporting a small number of formats, standards, and software will not meet the current practices of the majority of geoscientists. New CI serving this user community must therefore either expect users to change their current work practices or support heterogeneity. Given that changing existing work practices poses a high barrier to technology adoption (Kim & Crowston 2011), approaches such as brokering and translation tools may be more successful than standardization.

While it may appear that standards are not, in fact, becoming more widely used, this conclusion would likely not apply to CI under a narrower scope. The scientists interviewed purposely represented a broad spectrum of geosciences. Within narrower domains, the uptake of shared standards may be higher, and CI based on those standards more common. The existence and persistence of large community data portals, like IRIS for seismic data (IRIS 2019) and GBIF for biodiversity data (GBIF 2019), which are supported by specific standard data formats, indicate that standards-based systems can be successful within disciplinary communities of practice.

The variety of scientists' responses to the question 'what standard(s) do you use' demonstrates a lack of knowledge concerning standards in a substantial set of the interviewees: items like 'DOI', 'GPS' and 'NOAA' are not standards. Interviewees supplied information about data standards and formats in separate sections of the questionnaire, but based on the responses, it is clear that in certain cases interviewees view them as one and the same–that is, by selecting a data format, one has settled on a standard. This observation is consistent with comments heard that scientists were in search of standards or guidelines to follow when storing their data, in part because they were hoping that their data would be usable and accessible by other scientists interested in their research. Just as formats and standards were blended together somewhat, so too were formats and repositories. For example, an interviewee might say they use USGS formats. In fact, they may mean a USGS data repository, where various formats are employed in order to archive data.

An outreach and education effort aimed at increasing knowledge about the appropriate use and benefits of different standards and formats may be valuable to any CI effort seeking to foster standardization.

4.4. User scenario approach

This study employed semi-structured interviews, supported by a user scenario template, to assess geoscientists' priorities for CI, and guide the EarthCube cyberinfrastructure effort. Other methods exist for collecting end-user priorities for technical development projects, such as end user workshops with focus groups and self-administered questionnaires. We selected semi-structured interviews primarily because they allow for a high level of detail and completeness of coverage, with some degree of flexibility to allow for follow-up questions (Bernard 2012: 182–183). EarthCube had already engaged in a series of end-user workshops that brought together specialists in a specific geosciences domain, such as ocean ecology or paleogeosciences, to identify and report back their communities' scientific challenges and CI needs and desires. These provided valuable insight, but the feedback was often at a high level of generalization or focused on the scientific application more than the technical need. For example, a workshop report might state the need for CI to support testing of field-derived models, whereas the interview could probe into what the nature of that support is: Did the scientist require access to high-performance computing? What are the best practices on how to carry out model testing? How accessible is the data needed to test models more completely?

Overall, the workshop reports were useful for identifying the consensus priorities of a large community, while the individual interviews provided deeper detail and an ability to probe until the connections between science expectations and CI requirements was clear for a smaller number of people.

Self-administered questionnaires are another option for gathering user requirements. Arksey & Knight (1999) provide a thorough comparison between the qualitative interviews and self-administered questionnaires, including their strengths relative to richness of response, sample size, anonymity, cost in time, type of application, and more. Here, we reflect more specifically on the strengths and weaknesses experienced during the UCWG effort.

The UCWG originally distributed the template and asked scientists to fill it out themselves (a self-administered [SA] questionnaire approach.) This was not successful because the respondents often did

not fully understand what was being requested in each question or tended to skip over the more technical sections in favor of explaining their research program in more detail. In contrast, the interview method was more effective at exploring the stories and perspectives of respondents in greater depth. It also provided an opportunity to dialogue with a scientist and to clarify and adapt questions. In interviews, we were able to explain and give examples to clarify questions until they were understood, and circle back to technical details as needed to ensure completeness. The level of detail requested in the user scenario template was not appropriate for a self-reporting approach, which is consistent with research findings in similar projects (Bonney et al. 2009). Additionally, SA questionnaire responses were often too sparse overall or incomplete.

The clear downside of the interview method was the human resources needed: interviews required initial communications to explain the project and schedule the interview, 1–2 hours to conduct the interview, 1–2 hours to convert the raw interview notes into a finished document in the template format, and then a round of emails with the scientists asking them to review the document and provide any corrections. As such, we were able to reach a much smaller number of respondents than with a self-administered (SA) questionnaire. SA instruments can also be anonymous, though this was not a significant concern to the population we surveyed. Respondents were given the option of keeping their user scenarios confidential (i.e., used by UCWG in it summary reporting, but not published on its own), and no scientist requested confidentiality.

As a type of structured data collection tool, SA questionnaires limit the depth of information that can be collected. For this reason, questionnaires are often used to conduct high-level, exploratory research (i.e., to identify areas of interest for further study). In comparison, semi-structured interviews allow an interviewer to follow up on topics of interest during the interview itself, as well as to clarify responses.

If we were to conduct this effort again, we would make several adjustments to the approach. First, we would provide more pre-interview contact and preparation. Some respondents struggled to provide a user scenario, wanting instead to list CI desires without providing details and context; a better understanding of the scenario goals would have reduced this problem. We would also provide background on the technical details asked for and why we found them important in advance. The technical section (lists of data types, volumes, formats, tools used, etc.) was sometimes incompletely or inconsistently filled out, indicating it was difficult for some of the interviewees to respond to. In general, we noticed that respondents with more technical experience were more comfortable with the process, and those scientists with little familiarity with IT development found it more challenging. Second, we would adjust the template to reduce the emphasis on describing the complete scientific workflow in detail. This often led to lengthy discussions on parts of the science that had little relevance to CI. While the overall scientific context is critical to understand at a fairly detailed level, at some point, extensive details of methods that do not touch on any CI needs were not necessary to collect.

5. Conclusions

This paper reports on an effort to solicit feedback from a broad community of geoscientists using semi-structured interviews to collect user scenarios. The aim was to better understand the CI needs and priorities of the geosciences community. Overall, we found that respondents expressed concern about integrating existing CI into their research, and working with other scientists across systems independent of whether they already exist or are in planning, and have an interest in data collection and storage needs particular to their fields that makes translation across even the most elaborate CI challenging. More specifically, common areas of need include: better access to data, primarily with respect to more data being accessible online, but also creating better connectivity between related data across repositories to facilitate integration; tools to enable working with multiple formats; increased sharing of software and other code; guidance about best practices, protocols, and standards; and stable funding to support repositories, tools and other cyberinfrastructure beyond development and into long-term operations.

In addition, our study demonstrated the value utility of semi-structured, interview-based user scenarios as a mechanism to assess the CI needs and priorities of scientific communities. Our results demonstrate the strength of an open-ended approach in soliciting responses for use scenarios from this broad, diverse community: the level of detail acquired, the ability to draw out patterns across multiple interviews, and the often divergent opinions on CI needs that were not present during other methods utilized (focus groups, charettes, etc.).

These user scenarios, regardless of the method of acquisition, do not provide formal requirements in the software engineering sense: they cannot be handed to a developer as a set of instructions. However, they do provide expressions of user challenges that, in many cases, are sufficiently detailed to support

high-level requirement development. This was not done for this library of user scenarios because prioritization is a necessary next step. Not every challenge or limitation faced by the interviewed geoscientists can or should be addressed with new CI development. While the user scenarios highlight common concerns and challenges blocking scientific progress, it is for programs like EarthCube, as well as funding agencies, to determine priorities for new CI.

What we are able to recommend is greater use of diverse methods to tease out and better comprehend the disparate priorities and needs expressed by scientists. It may behoove CI designers and implementers to consider user scenarios based on semi-structured interviews like those carried out for this paper, given the type of data elicited relative to other methods typically employed. The results of this work show that many patterns emerge across scenarios, all of which may be relevant to those in the field of CI.

Acknowledgements

This work was carried out through the Use Case Working Group under the EarthCube Technology and Architecture Committee. We appreciate the valuable input and insight of the Use Case Working Group former chair, Danie Kinkade, working group members (Yolanda Gil, Tanu Malik, Chris Mattman, Scott Peckham, and Plato Smith), and the guidance of Jay Pearlman as EarthCube Technology and Architecture Committee co-chair. This work was supported by the National Science Foundation, under award ICER 1343813.

Competing Interests

Lisa Kempler is employed at MathWorks, which makes Matlab, one of the software tools used by the geosciences community. All other authors have no competing interests.

Author Contributions

Karen Stocks and Lisa Kempler conceived of the project, developed the use case template, conducted initial test interviews, analyzed and synthesized the data, and wrote the initial manuscript draft. Stocks also acted as overall project manager. Sam Schramski and Arika Virapongse carried out the majority of the use case interviews, and contributed to the writing of the manuscript, with Schramski leading paper revision and submission.

Author Information

Karen Stocks is the Director of the Geological Data Center at Scripps Institution of Oceanography, where she specializes in the documentation, discovery, access, integration, and curation of oceanographic data. Her expertise includes information systems for vessel-based sensors, scientific ocean drilling, biodiversity and biogeography, metagenomics, and ocean observing systems.

Arika Virapongse is a Research Scholar at the Ronin Institute and Principal Scientist at Middle Path EcoSolutions. Arika works on data integration and application for social-ecological systems, community development, and socioeconomic value assessment.

Sam Schramski is a Research Associate affiliated with the Center for the Analysis of Social-Ecological Landscapes (CASEL) at Indiana University. He is currently working on data science studies at the intersection of geography, anthropology and environmental science.

Lisa Kempler is the MATLAB Community Strategist at MathWorks, and works on geoscience community development, data analytics, and scientific software product management. Lisa holds a Bachelors in Computer Science from Brown University and a Masters in Computer Information Systems from Boston University. She serves as Co-chair of the EarthCube Use Case Working Group.

References

American Geophysical Union. 2018. Index Terms–Publications. Available at http://publications.agu.org/author-resource-center/index-terms/ [Last accessed 21 December 2018].

Arksey, M and **Knight, PT.** 1999. *Interviewing for Social Scientists: An Introductory Resource with Examples.* London, UK: SAGE Publications. DOI: https://doi.org/10.4135/9781849209335

Asmi, A, Brus, M and **Sorvari, S.** 2017. Community-Driven Efforts for Joint Development of Environmental Research Infrastructures. In: Chabbi, A and Loescher, HW. *Terrestrial Ecosystem Research Infrastructures: Challenges, New developments and Perspectives.* Boco Raton, FL: CRC press. DOI: https://doi.org/10.1201/9781315368252-18

Assistant Secretary for Public Affairs. 2013. Scenarios|Usability.gov. 29 May 2013. Available at https://www.usability.gov/how-to-and-tools/methods/scenarios.html. [Last accessed 21 December 2018].

Bernard, HR. 2012. *Social research methods: qualitative and quantitative approaches*. Los Angeles, CA: Sage.

Bonney, R, Cooper, CB, Dickinson, J, Kelling, S, Phillips, T, Rosenberg, KV and **Shirk, J.** 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, 59: 977–984. DOI: https://doi.org/10.1525/bio.2009.59.11.9

Carroll, JM. 2000. *Making use: scenario-based design of human-computer interactions*. Cambridge, MA: MIT Press.

Cohn, M. 2009. *Succeeding with Agile: Software Development Using Scrum*, vol. 1. Boston, MA: Addison-Wesley Professional.

Cutcher-Gershenfeld, J, Baker, KS, Berente, N, Carter, DR, DeChurch, LA, Flint, CC, Gershenfeld, G, Haberman, M, King, JL, Kirkpatrick, C, Knight, E, Lawrence, B, Lewis, S, Lenhardt, WC, Lopez, P, Mayernik, MS, McElroy, C, Mittleman, B, Nichol, V, Nolan, M, Shin, N, Thompson, CA, Winter, S and **Zaslavsky, Z.** 2016. Build It, But Will They Come? A Geoscience Cyberinfrastructure Baseline Analysis. *Data Science Journal*, 15(8). DOI: https://doi.org/10.5334/dsj-2016-008

Dooley, R, Milfeld, K, Guiang, C, Pamidighantam, S and **Allen, G.** 2006. From Proposal to Production: Lessons Learned Developing the Computational Chemistry Grid Cyberinfrastructure. *Journal of Grid Computing*, 4: 195–208. DOI: https://doi.org/10.1007/s10723-006-9043-7

EarthCube. 2018. Homepage. Available at: https://www.earthcube.org. [Last accessed 15 January 2019].

Finholt, TA and **Birnholtz, JP.** 2006. If we build it, will they come: the cultural challenges of cyberinfrastructure development. In: Banbridge, W and Rocco, MC. *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society*. Netherlands: Springer. pp. 89–101. DOI: https://doi.org/10.1007/1-4020-4107-1_7

GBIF (Global Biodiversity Information Facility). 2019. Available at: https://www.gbif.org. [Last accessed 15 January 2019].

Geertz, C. 1973. Thick Description: Toward and Interpretive Theory of Culture. In: *The Interpretation of Cultures*. New York, NY: Basic Books. pp. 3–32.

IRIS (Incorporated Research Institutions for Seismology). 2019. Available at: www.iris.edu. [Last accessed 15 January 2019].

Kee, KF and **Browning, LD.** 2010. The Dialectical Tensions in the Funding Infrastructure of Cyberinfrastructure. *Computer Supported Cooperative Work (CSCW)*, 19: 283–308. DOI: https://doi.org/10.1007/s10606-010-9116-9

Kim, Y and **Crowston, K.** 2011. Technology adoption and use theory review for studying scientists' continued use of cyber-infrastructure. *Proceedings of the American Society for Information Science and Technology*, 48: 1–10. DOI: https://doi.org/10.1002/meet.2011.14504801197

NIST Big Data Working Group. 2017. Use Cases and Requirements. 11 January 2017. Available at https://bigdatawg.nist.gov/usecases.php [Last accessed 21 December 2018].

Open Access to Ocean Data. 2018. AODN Portal v4.38.37. Available at: https://portal.aodn.org.au/. [Last accessed: 15 January 2019].

Ribes, D and **Lee, CP.** 2010. Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. *Computer Supported Cooperative Work*, 3/4: 231–244. DOI: https://doi.org/10.1007/s10606-010-9120-0

Short, JC, Broberg, JC, Cogliser, CC and **Brigham, KH.** 2010. Construct Validation Using Computer-Aided Text Analysis (CATA): An Illustration Using Entrepreneurial Orientation, *Organizational Research Methods*, 13: 320–347. DOI: https://doi.org/10.1177/1094428109335949

Stocks, KI, Schramski, S, Virapongse, A and **Kempler, L.** 2019. EarthCube User Scenario Collection. *UC San Diego Library Digital Collections*. DOI: https://doi.org/10.6075/J0WQ024C

Tetherless World Constellation. 2014. *Use Cases*. Available at https://tw.rpi.edu/web/UseCases [Last accessed 21 December 2018].

Yin, RK. 2017. Collecting Case Study Evidence: The Principles You Should Follow in Working With Six Sources of Evidence. In: *Case Study Research and Applications: Design and Methods*. Los Angeles, CA: SAGE Publications. pp. 111–164.

How to cite this article: Stocks, KI, Schramski, S, Virapongse, A and Kempler, L. 2019. Geoscientists' Perspectives on Cyberinfrastructure Needs: A Collection of User Scenarios. *Data Science Journal,* 18: 21, pp. 1–15. DOI: https://doi.org/10.5334/dsj-2019-021

Submitted: 19 January 2019 **Accepted:** 23 May 2019 **Published:** 18 June 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS