DATA SCIENCE JOURNAL

Cutcher-Gershenfeld, J et al 2016 Build It, But Will They Come? A Geoscience Cyberinfrastructure Baseline Analysis. *Data Science Journal,* 15: 8, pp. 1–14, DOI: http://dx.doi.org/10.5334/dsj-2016-008

RESEARCH PAPER

Build It, But Will They Come? A Geoscience Cyberinfrastructure Baseline Analysis

By the Stakeholder Alignment Collaborative[1]: Joel Cutcher-Gershenfeld[2], Karen S. Baker[3], Nicholas Berente[4], Dorothy R. Carter[5], Leslie A. DeChurch[6], Courtney C. Flint[7], Gabriel Gershenfeld[8], Michael Haberman[9], John Leslie King[10], Christine Kirkpatrick[11], Eric Knight[12], Barbara Lawrence[13], Spenser Lewis[14], W. Christopher Lenhardt[15], Pablo Lopez[16], Matthew S. Mayernik[17], Charles McElroy[18], Barbara Mittleman[19], Victor Nichol[20], Mark Nolan[21], Namchul Shin[22], Cheryl A. Thompson[23], Susan Winter[24] and Ilya Zaslavsky[25]

[1] Members of the Stakeholder Alignment Collaborative involved in this article

[2] Heller School for Social Policy and Management, Brandeis University, 415 south street, room 202, Waltha. Ma 02453, 2700

[3] Graduate School of Information Sciences, University of Illinois at Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820, US

[4] Terry College of Business, University of Georgia, Athens GA 30602, US

[5] Psychology Department, University of Georgia, 125 Baldwin Street, Athens, GA 30602, US

[6] School of Psychology, Georgia Institute of Technology, 654 Cherry St. Atlanta, GA 30332, US

[7] Dept of Sociology, Social Work & Anthropology, Utah State University, 0730 Old Main Hill, Logan, UT, USA 84322, US

[8] Gabe Gershenfeld, Los Angeles Dodgers, 1000 Elysian Park Ave., Los Angeles, CA 90012, US

[9] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 1205 W. Clark St., Urbana, Illinois, US

[10] School of Information, University of Michigan, 105 S. State St. 3447, Ann Arbor MI 48109, US

[11] San Diego Supercomputer Center, University of California San Diego, La Jolla, California 92093-0505, US

[12] Discipline of Work & Organisational Studies, University of Sydney Business School, Abercrombie Building (H70), The University of Sydney, NSW 2006, US

[13] Anderson Graduate School of Management, University of California Los Angeles, Los Angeles, California 90095-1481, US

[14] Draper labs 555 Technology Square, Cambridge, MA 02139, US

[15] Renaissance Computing Institute University of North Carolina, 100 Europa Drive, Suite 540, Chapel Hill, NC 27517, US

[16] Department of Statistics, University of Illinois, Urbana-Champaign, 1205 W. Clark St., Urbana, IL 61801, US

[17] National Center for Atmospheric Research, University Corporation for Atmospheric Research, Boulder, CO, US

[18] Department of Design and Innovation, Weatherhead School of Management, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, US

[19] WayMark Systems, Inc., Enterprice works, 60 Hazelwood Drive, Champaign, IL 61820

[20] University of Illinois at Urbana-Champaign, EarthCube project; 130 West 1st Street Niles, OH 44446, US

[21] Graduate School of Information Sciences, University of Illinois at Urbana-Champaign, 501 E. Daniel St., Champaign, Illinois, USA 61821, US

[22] Seidenberg School of Computer Science and Information Systems, Pace University, 163 William Street, New York, NY 10038, US

[23] Graduate School of Information Sciences, University of Illinois at Urbana-Champaign, 501 E. Daniel St., Champaign, Illinois, USA 61821, US

[24] College of Information Studies, University of Maryland, US

[25] San Diego Supercomputing Center, University of California, San Diego, US

Corresponding author: Joel Cutcher-Gershenfeld (joelcg@brandeis.edu)

Understanding the earth as a system requires integrating many forms of data from multiple fields. Builders and funders of the cyberinfrastructure designed to enable open data sharing in the geosciences risk a key failure mode: What if geoscientists do not use the cyberinfrastructure to share, discover and reuse data? In this study, we report a baseline assessment of engagement

with the NSF EarthCube initiative, an open cyberinfrastructure effort for the geosciences. We find scientists perceive the need for cross-disciplinary engagement and engage where there is organizational or institutional support. However, we also find a possibly imbalanced involvement between cyber and geoscience communities at the outset, with the former showing more interest than the latter. This analysis highlights the importance of examining fields and disciplines as stakeholders to investments in the cyberinfrastructure supporting science.

Keywords: Curation; Cyberinfrastructure; EarthCube; Fields; Disciplines; Geoscience; Infrastructure; Network effects; Open data; Reuse; Stakeholder Alignment

1 Introduction

Builders and funders of data infrastructure always bear the risk that subsequent use and impact will fall short of expectations, calling into question what are often substantial up-front investments. In this article we use survey results of geoscientists and cyberinfrastructure experts to assess potential barriers for the U.S. National Science Foundation's EarthCube initiative, which is advancing the cyberinfrastructure for geoscience. The goal of EarthCube is to serve the geosciences by integrating unique data sets, isolated repositories, separate models, and relevant software through functional tools, thus allowing discovery and reuse of diverse geoscience data. Expanding open sharing and data discovery supports replicability and integrity in science, enables new frontiers of research within and across disciplines, and informs policy decisions. Such integration will likely accelerate progress on grand challenges, such as global climate change, severe weather prediction, natural resource discovery, and, ultimately, understanding the earth as a system. Three potential barriers are assessed: lack of agreement on the need for the infrastructure, lack of support for engaging with the infrastructure, and a lack of balance in engagement of diverse communities.

Because the EarthCube initiative is substantial in scope and vision, a decade or more will be needed for development and operation. EarthCube's success depends on the engagement of all geoscience fields (we use this term to indicate disciplines as well). The community needs to develop a culture in which physical samples, data on physical samples, streaming data from sensors, visual images, data models, and other aspects of geoscience data are curated, shared, and reused. Indeed, there is public policy support for this sharing and reuse (OSTP Public Access Memo, 2013). EarthCube's success further depends on expanding workflows to provide appropriate information on data provenance, systems tracking data reuse and providing credit for reuse (comparable to citation counts), universities valuing such behaviors to a much greater degree in promotion and tenure decisions, and appropriate industry partnerships. In addition to the technical infrastructure, data sharing success depends on increasing levels of trust and cooperation across diverse geoscience fields. EarthCube is one of a number of data sharing initiatives emerging to facilitate work across diverse disciplines and public engagement in science. Some of these are long-standing, such as CODATA, and others are newer, such as CyVerse (formerly iPlant), the iSamples initiative, the Materials Genome Initiative (MGI), the National Data Service (NDS), and the Research Data Alliance (RDA).

Although the potential value of EarthCube is clear, success is not assured. In 2012, one year after the formal launch, a second launch conference, termed a "charrette," engaged over 200 experts in road-mapping exercises. However, the participant mix included many more cyberinfrastructure builders than geoscience end-users. At the conclusion of the session a dozen NSF leaders and session facilitators met and identified seven potential failure modes of EarthCube.[1] One of these was viewed as particularly problematic. It was stated as follows: "The 'build it and they will come' mindset – in which users don't show up, data are not shared." In this article we present survey data exploring three potential barriers relevant to this possible failure mode. These potential barriers are I. Geoscientists do not perceive the need for change;

[1] Other potential failure modes mentioned at the time included:
- Unrealistic or misaligned expectations among people presently involved in EarthCube
- Not valuing what presently exists – current cyber/geo science efforts and initiatives that represent parts of the EarthCube vision
- Not advancing the frontier in transformative ways relative to what presently exists – only automating the current state
- Not engaging the 200,000+ geoscience and cyber stakeholders not presently involved in EarthCube
- Not anticipating the needs of the next generation of geoscience and cyber stakeholders (today's doctoral students and post docs, as well as the generation behind them)
- "Unk Unk" – additional unknown unknowns including transformational changes in the technology, disruptive shifts in the policy arena, etc.

II. Geoscientists do not have support for engaging in change initiatives; and III. The disciplinary mix of engaged users is not appropriately balanced. Importantly, once addressed, the barriers can become drivers, helping to advance EarthCube. These barriers/drivers are likely common across similar projects in all disciplines involving what is sometimes termed big data.

2 Build it and they will come

The epigram "build it and they will come" or the negative formulation "build it, but they don't come" has entered common parlance in domains as diverse as information systems design (Markus & Keil, 1994); marketing management (Henrix, 1999); medical services (Tintinalli, 2008), and organizational learning (Yuan, et. al., 2010). In these and other cases, the concern is that target audiences do not come – in contrast to the popularization of the concept in the novel *Shoeless Joe* (W.P. Kinsella, 1982) and the 1989 movie adaptation, "Field of Dreams," where faith was rewarded."[2]

In the case of EarthCube, the builders include cyberinfrastructure experts, data managers, software engineers, data scientists, and others. The "they" who might or might not come are geoscientists in a broad array of fields. Supporting and conducting research on users as diverse as these is difficult since it involves reference groups that are not easily specified (Lawrence, 2006), though they may be commonly recognizable.[3]

Divides between technologists and scientists have been documented in a wide range of geoscience settings (Mayernik, Wallis, & Borgman, 2013; Jackson & Buyuktur, 2014; Finholt and Birnholtz, 2006; Ribes and Finholt, 2008). These divides range from challenges in developing effective collaborative structures and aligning incentives, to difficulties in establishing leadership roles and determining who the relevant "community" actually is for a given cyberinfrastructure initiative. For example, Finholt and Birnholtz (2006) note that domain scientists likely will be more successful in leading cyberinfrastructure initiatives than technical experts, since they can marshal participation by their scientific colleagues. In "build it and they will come," the "build it" part is more certain (and the costs are paid for). Whether "they will come" is more uncertain. At a time when broader impacts are expected of science and of the investments that support science, both parts of the statement are essential. Our large-scale survey was designed to identify key factors in the success of the NSF EarthCube initiative. The idea of barriers and drivers to system change was initially advanced by Kurt Lewin (1939) who noted that barriers, once addressed, often become drivers. He also noted that reducing barriers produced greater leverage in the "force field" than pushing harder on the drivers.

Further underlying the idea of "build it and they will come" is the expectation of value through what are termed "network effects" (Katz and Shapiro, 1994). If people from a sufficiently broad network share and discover data, the system will work. If not, the system is at risk of not having a self-sustaining critical mass of users.

3 Background on EarthCube

The NSF initiated EarthCube in 2009 and charged it with creating the cyberinfrastructure for advancing regular science and addressing grand challenges in the geosciences that involve understanding the earth as a system. In October, 2011, EarthCube conducted its first charrette, which drew over 150 participants. In February, 2012, the NSF commissioned what was termed a "stakeholder-alignment survey" for the second charrette in June, 2012. The survey was designed to indicate points of alignment or misalignment among geo and cyber stakeholders, drawing on methods developed under a prior NSF research grant (NSF-VOSS EAGER 0956472). Although the second charrette drew even more attendees, the high proportion of computer and data experts and the initial survey data both pointed to a need for increased engagement of geoscientists. In response, the NSF made a strategic pivot to solicit proposals for geoscientist-led disciplinary workshops on data sharing. Before each workshop participants were invited to complete the stakeholder alignment survey.

---

[2] In the novel and the movie, the expression is "Build it and he will come," referring to the ghost of Shoeless Joe Jackson. This more positive formulation has deeper roots. For example, Ralph Waldo Emerson observed: "Build a better mousetrap, and the world will beat a path to your door." Of course, in the Bible, Noah was instructed to build an ark on the faith that the animals and, ultimately, salvation would come.

[3] It also involves strongly felt issues of identity. For example, among 1,569 responses to our survey, there were over 700 unique responses when was asked about people's specific expertise, including specific specialties such as: Air Sea Interaction, Basalt geochemistry, Biodiversity Information Networks, Carbonate Stratigraphy, Coastal Geomorphology, Computational Geodynamics, Cryosphere-Climate Interaction, Ensemble data assimilation, Geomicrobiology, Heliophysic, Isotope Geochemistry, "It's complicated," Magnetospheric Physics, Mesoscale Meteorology, Multibeam Bathymetric Data, Paleoceanography, Permafrost Geophysics, Riverine carbon and nutrient biogeochemistry, Satellite gravity and altimetry data processing, and Thermospheric Physics.

A further impetus for EarthCube emerged on February 22, 2013, when the U.S. Office of Science and Technology Policy (OSTP) directed federal executive offices and agencies to make the direct results of federally funded scientific research available. The directive stated, in part:

> *The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law . . . the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data. . . . These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.*

This was followed by similar directives on January 1, 2013 from the Australian Research Council and on May 24, 2013, from the UK Research Councils. Most recently, following the end-user workshops and the data collection, EarthCube has developed a system of governance for community engagement and cyberinfrastructure development. Here, we present baseline data collected in advance of the governance and more recent activities.

4 Data Sources

The stakeholder-alignment surveys were designed to enable evidence-based decision making for EarthCube and to provide a baseline for later surveys every 12–18 months. The initial survey was developed in March and April 2012, approved by the University of Illinois Institutional Review Board, and administered via a web-based URL. Questions asked cyberinfrastructure developers and geoscientists for their views of the EarthCube concept and indications of their involvement. Comparing responses across groups indicates the current degree of alignment among builders and potential users and helps identify the barriers to increased alignment that can jeopardize the EarthCube initiative. In this case, scientific fields and professional domains are the stakeholders.

Initially 167 out of 900 website registrants completed the survey. Respondents included a large group of cyberinfrastructure builders. Subsequent outreach using lists from NSF data centers yielded an additional 578 responses, out of approximately 10,000 requests, with a higher proportion of geoscientists (with 1 unspecified).

Beginning in the fall of 2012, throughout 2013, and into 2014, participants at 23 disciplinary workshops also completed the survey. In January 2013, the survey was revised (after the first five workshops). Responses from all workshops were received from 824 of the 1,828 invitees (spanning versions 1 and 2). We used the entire sample of 1,569 respondents in our analysis (167 + 578 + 824). The workshop response rates are indicated in **Table 1**. The invitees to these workshops were a combination of thought-leaders identified by NSF program officers, individuals contacted through professional associations and societies, and early career scientists. The fields featured at the workshops were based on competitive workshop proposals submitted to the NSF.

We use descriptive statistics, bar charts, a new data visualization format, graphing, and OLS regression modeling to assess potential barriers. In the bar charts and the regression analyses we use the fields and disciplines that respondents selected from a taxonomy included in the survey (different than the workshop titles listed above). Many survey questions use a 16-point scale, which is transposed to a scale from zero to one for ease of interpretation. Both survey instruments included voluntary consent language and received IRB approval at the University of Illinois, Urbana-Champaign. The data is available as DOI: http://dx.doi.org/10.1594/IEDA/100535 (Title: EarthCube Stakeholder Survey Data; Date Available: 2015-04-15; URL: http://dx.doi.org/10.1594/IEDA/100535). In addition, an interactive tool for exploring the data, allowing for two-way charting and other forms of exploratory analysis, is available at: http://maxim.ucsd.edu/suave/survey1544.html.

The stakeholder alignment approach is modeled on previous uses of systematic stakeholder data to inform complex multi-stakeholder initiatives including a 2005 report to the U.S. Congress on aircraft noise and emissions (Waitz, et. al.) and NSF-funded studies of the BioMarkers Consortium, the U.S. Fab Lab Network, and local community green energy (NSF-VOSS EAGER 0956472). Similar stakeholder alignment surveys have been conducted for the National Data Service, the i-Samples initiative, and other complex multi-stakeholder projects. We define stakeholder alignment as "the dynamic process by which interdependent stakeholders orient and connect to advance separate and shared interests." In this article, scientific fields and related technical domains are all treated as stakeholders to the cyberinfrastructure supporting open data for science. The needed alignment is not a fixed end point, but an ongoing process involving many shared and separate interests.

Version 1 of the survey instrument:		
1. Early Career	24.7% (n =3 7 of 150)	Oct. 17–18, 2012
2. Structure and Tectonics	70.5% (n = 24 of 34)	Nov. 19–20, 2012
3. EarthScope	31.9% (n = 22 of 69)	Nov. 29–30, 2012
4. Experimental Stratigraphy	42.9% (n = 21 of 49)	Dec. 11–12, 2012
5. Atmospheric Modeling / Data Assimilation and Ensemble Prediction	31.2% (n = 29 of 74)	Dec. 19, 2012
Version 2 of the survey instrument:		
6. OGC	28.0% (n = 14 of 50)	Jan. 13, 2013
7. Critical Zone	28.3% (n = 39 of 138)	Jan. 21–23, 2013
8. Hydrology / Envisioning a Digital Crust	48.9% (n = 23 of 47)	Jan. 29–31, 2013
9. Paleogeoscience	50.6% (n = 40 of 79)	Feb. 3–5, 2013
10. Education & Workforce Training	57.9% (n = 33 of 57)	Mar. 3–5, 2013
11. Petrology & Geochemistry	71.1% (n = 59 of 83)	Mar. 6–7, 2013
12. Sedimentary Geology	55.6% (n = 50 of 90)	Mar. 25–27, 2013
13. Community Geodynamic Modeling	46.4% (n = 45 of 97)	Apr. 22–24, 2013
14. Integrating Inland Waters, Geochemistry, Biogeochem and Fluvial Sedimentology Communities	39.0% (n = 46 of 118)	Apr. 24–26, 2013
15. Deep Sea Floor Processes and Dynamics	49.2% (n = 29 of 59)	June 5–6, 2013
16. Real-Time Data	23.4% (n = 25 of 107)	June 17–18, 2013
17. Ocean 'Omics	71.2% (n = 42 of 59)	Aug. 21–23, 2013
18. Coral Reef Systems (two workshops)	91.7% (n = 44 of 48)	Sept. 18–19/ Oct. 23–24, 2013
19. Geochronology	44.6% (n = 66 of 148)	Oct. 1–3, 2013
20. Ocean Ecosystem Dynamics	45.0% (n = 36 of 80)	Oct. 7–8, 2013
21. Clouds and Aerosols	63.9% (n = 39 of 61)	Oct. 21–22, 2013
22. Rock Deformation and Mineral Physics	44.3% (n = 37 of 79)	Nov. 12–14, 2013
23. Marine Seismic	46.2% (n = 24 of 52)	Dec. 11–12, 2014

Table 1: Response Rates and Timing of NSF EarthCube Disciplinary Domain Workshops (n=824 of 1,828).

5 Potential Barrier I: Geoscientists do not perceive the need for change

To assess the need for change, respondents were asked four key questions about both the importance and the ease of finding, accessing, and integrating data, models, and software:

- *How IMPORTANT is it for you to find, access, and/or integrate multiple datasets, models, and/or software (e.g. visualization tools, middleware, etc.) in your field or discipline?*
- *How EASY is it for you to find, access, and/or integrate multiple datasets, models, and/or software (e.g. visualization tools, middleware, etc.) in your field or discipline?*
- *How IMPORTANT is it for you to find, access, and/or integrate multiple datasets, models, and/or software (e.g. visualization tools, middleware, etc.) that span different fields or disciplines?*
- *How EASY is it for you to find, access, and/or integrate multiple datasets, models, and/or software (e.g. visualization tools, middleware, etc.) that span different fields or disciplines?*

The mean and standard deviation for importance within fields is 0.87 +/– .20, while the result for ease within fields is 0.41 +/– .25. This same gap is found with respect to perceived importance across fields (0.76 +/– .26) as compared with the perceived ease of doing so (0.30 +/– .23).

For the EarthCube participants to better view the data, we used a newly developed format, termed a z-flower. Each respondent is assigned a color-coded hexagon with shades of green for positive responses

(darker is more positive), shades of yellow for neutral responses, and shades of red for negative responses, and the white hexagons represent those who didn't answer the question. The responses are then tiled in a spiral with those close to the mean in the middle (indicating the central tendency) and those farthest from the mean on the outside (the outliers). This heightens the contrast between importance and ease (**Figure 1**).

The combined responses (**Figure 2**) are similar and consistent across fields. Overall, the consistent gap between ease and importance, which was greater than 0.5 across nearly all fields, supports the perceived need for the EarthCube initiative. Also the gap across fields is slightly but consistently larger than within them. These data quantify the frustrations of researchers in sharing or obtaining data in their fields and across.

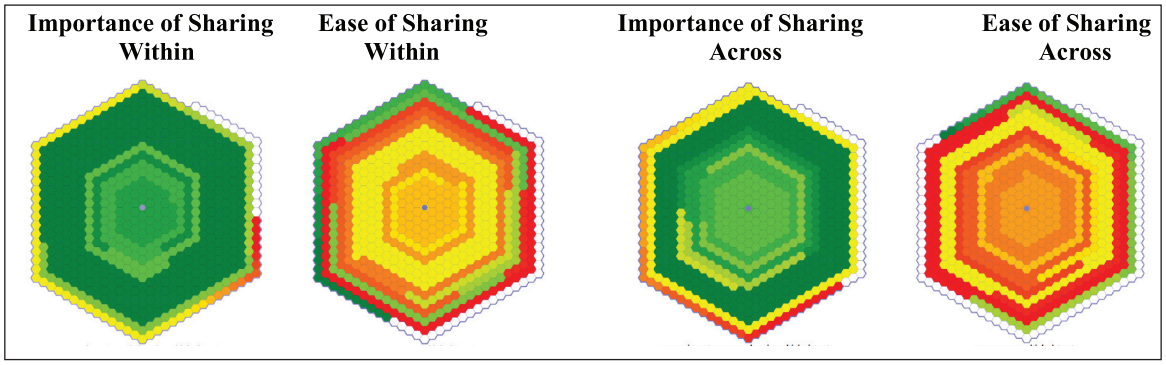

Figure 1: Visual Representation of Importance and Ease of Sharing Data Within and Across Fields.
Note: The contrast between importance and ease can be seen within and across fields. There are few cases where it is not important and a few cases where it is seen as easy (the outliers on the outside edges of each z-flower), but these are small compared to the central tendencies. Each small hexagon represents a respondent. Shades of green signal positive views; shades of yellow, neutral views; shades of red, negative views. Responses are tiled from the middle, which is the mean, in a spiral outward above and below the mean, so the middle is the central tendency and outliers are on the outside. Missing responses are white.

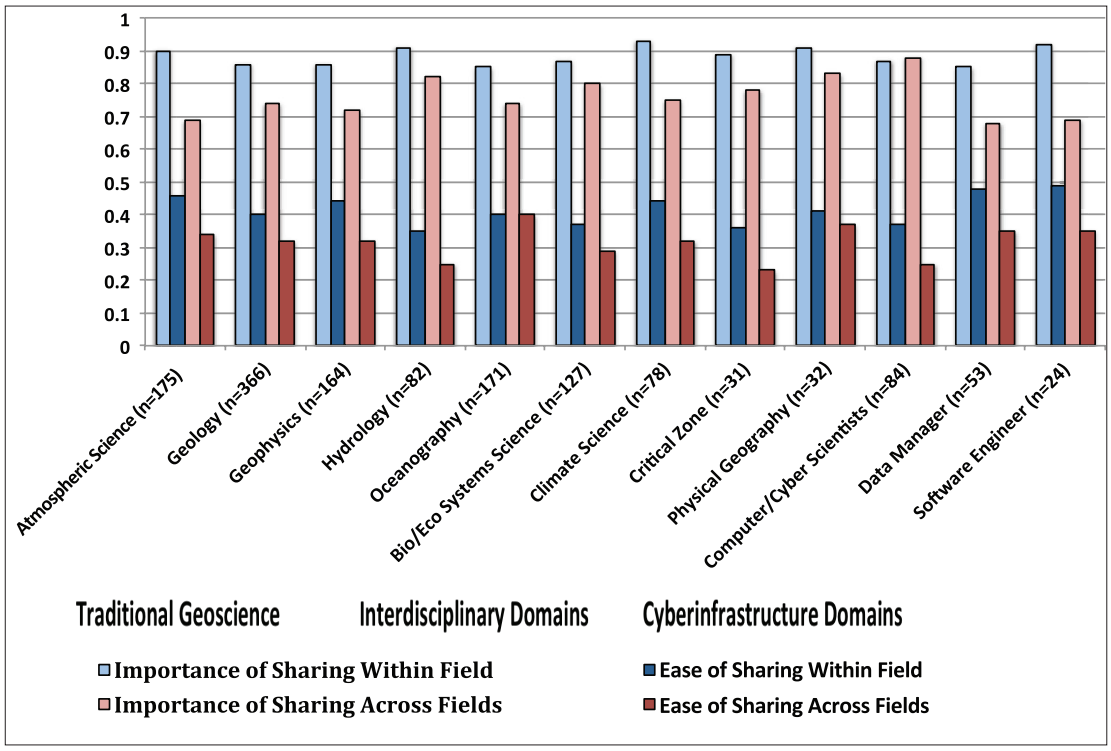

Figure 2: Ease and Importance of Sharing Data, Tools, and Models: Within and Across Fields.
Note: The gaps between importance and ease within and across fields are remarkably similar in these diverse fields. Bar charts are used rather than z-flowers to compare means across multiple fields (in addition to illustrating the full diversity of responses.

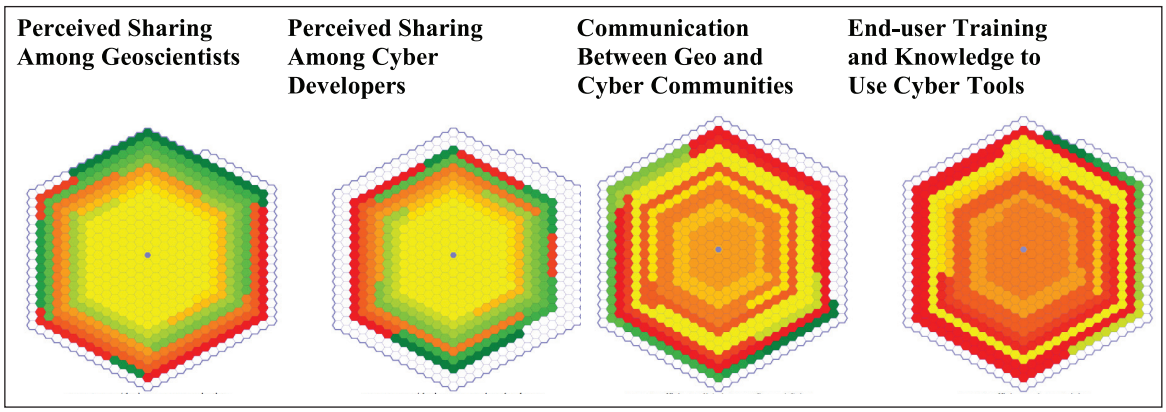

Figure 3: Visual Representation of Cooperation Between Geo and Cyber and End-User Knowledge and Training in the Cyberinfrastructure.

Note: While there are some very positive and very negative outliers in perceptions of sharing among geoscientists and among cyber developers (the first and second z-flowers), the dominant view is at the midpoint on the scale. Note that many respondents indicate no response when asked about cooperation among cyber developers. The communication between geo and cyber is seen negatively and the responses are even more negative when asked about end-user training and knowledge on the use of cyber tools. Each small hexagon represents a respondent. (see note on figure 1 for additional explanation on reading these figures.)

Other responses reinforced the above findings. For example, respondents were primarily neutral or negative, with a few positive "bright spots" to the statements: "There is currently a high degree of sharing of data, models, and software among geoscientists" and "There is currently a high degree of sharing of software, middleware and hardware among those developing and supporting cyberinfrastructure for the geosciences." Respondents held more negative perceptions to the statements: "There is currently sufficient communication and collaboration between geoscientists and those who develop cyberinfrastructure tools and approaches to advance the geosciences" and "There is currently sufficient geoscience end-user knowledge and training so they can effectively use the present suite of cyberinfrastructure tools and train their students/colleagues in its use" as is illustrated by **Figure 3**.

Clearly, getting access to others' data is seen as important, but not easy. Respondents perceive a need for change, which can serve as a driver (rather than a barrier) for EarthCube. This is promising for EarthCube.

6 Potential Barrier II: Geoscientists do not have support for engaging in change initiatives

A second potential barrier to the success of EarthCube arises if geoscientists perceive that their employing organization, such as a university or data center, and professional colleagues do not support engagement with EarthCube. Two questions in the survey addressed these forms of support:

- *My employer/organization will most likely value and reward any efforts I make in the shaping and development of EarthCube.*
- *Any contributions I might make to the shaping and development of EarthCube will likely be recognized and valued by colleagues in my field/discipline.*

We combined these two items into a single EarthCube support scale (Cronbach's alpha = 0.79).

The results of the initial survey suggested that the lack of such support could indeed be a barrier to EarthCube's success. Thus, two additional questions were added to the survey in January 2013 to see if this was specific to EarthCube or a more general barrier around interdisciplinary science:

- *My employer/organization will value and reward efforts I make in bridging across fields and disciplines.*
- *Efforts that I make to bridge across fields and disciplines will most likely be recognized and highly valued by colleagues in my field/discipline.*

We combined these two items into a single interdisciplinary support scale (Cronbach's alpha = 0.75).

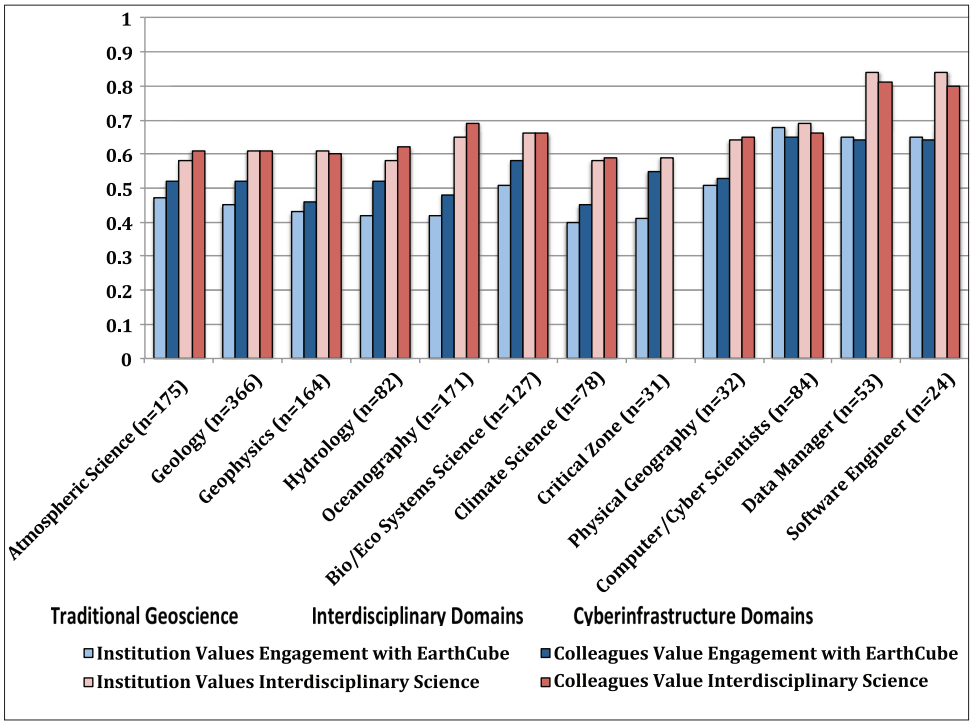

Figure 4: Support from Employer/Institution and from Colleagues for Engagement with EarthCube and for Interdisciplinary Science.

Note: The general support for interdisciplinary science is greater than the specific support for engagement with EarthCube. The levels of support are higher among the cyber domains. The number of respondents for the second set of interdisciplinary questions is about half the total listed number of respondents since that question was added part way through data collection.

The responses (**Figure 4**) indicate only moderate perceived support on average for engagement from the respondent's employing organization (0.48 +/– .31) and colleagues (0.53 +/– .27) across diverse fields. Moreover, the standard deviations are relatively large, indicating considerable variation within the geosciences. Note that these responses do not include the first version of the survey in order to compare the responses with the next question discussed (on support for interdisciplinary science), which was only added in the second survey.

Respondents' support for engagement in interdisciplinary science is more positive on average at 0.64 +/– .24 than support for engagement with EarthCube. This is consistent with their perceived support from employing organization and 0.64 +/– .25 from colleagues.

In general, driving change is more difficult when there is a high degree of variability (Deming, 1986), suggesting that initial engagement of researchers in EarthCube will be constrained by the lack of strong and consistent support from their home organizations or colleagues. As a baseline assessment of this potential barrier, the results are mixed. Some end-users report support from their home organization and colleagues, but in most cases support is not strong. This suggests that institutional infrastructure will need attention for EarthCube or any cyberinfrastructure initiative to achieve full success. In a separate working paper (Stakeholder Alignment Collaborative, 2015), we focus more deeply on the underlying theory of internal support needed within fields to engage in multi-stakeholder collaboration.

7 Potential Barrier III: The disciplinary mix of engaged users is not appropriately balanced

To assess the degree to which geoscientists are engaged with EarthCube, we use a bivariate figure and multivariate analysis. Respondents' degree of engagement in the EarthCube initiative, using a 6-point ordinal scale, is the anchor for the bivariate analysis and the dependent variable for the multivariate analysis:

1 = I have heard of EarthCube (10.2%)
2 = Aware of EarthCube, but no engagement (33.7%)

3 = Visited the website (16.5%)
4 = Participated in discussions (17.4%)
5 = Actively involved (16.7%)
6 = Leadership role (5.1%)

Increasing values represent increasing degrees of engagement (0.4% non-responses are excluded). We don't know the degree to which engagement in EarthCube is representative of other open data or data sharing initiatives. However, increased sharing and reuse of data will involve a mix of individual actions and collective initiatives that involve increasing degrees of engagement. Because EarthCube was still in early stages of formation at the time of data collection, the overall averages indicate that most participants were just at the level of awareness. The bivariate analysis in **Figure 5** shows engagement with EarthCube on one axis and support for interdisciplinary science from the respondent's employing organization or institution on the other axis (a scale from 0–1). Only the relevant portions of each scale are used in this figure and the means are marked as mid-points on each axis. Contour lines are added in the open spaces between clusters of fields and disciplines; they are added for illustrative purposes.

Figure 5 suggests that cyberinfrastructure builders (above the first contour line) have more support and engagement with EarthCube than end-users. This is reasonable: the concept "build it and they will come" states up-front that it will be built. Those who build the infrastructure know they will be paid. Whether "they" – the end-users – will come is an open question. The risk is not in building something. Often it is impossible to tell whether "they will come" until something is built. The risk is that the built artifact will not be what is needed. The second contour points to fields with the greatest initial risk.

Multiple regression was used to test a multivariate model in which the control variables are experience, gender, and international or domestic home organization. To assess the balance of engagement among end-users across fields we focus on a set of traditional geoscience disciplines (atmospheric/space weather science, geology, geophysics, hydrology, and oceanography) and a set of interdisciplinary scientific domains (bio/ecosystem science, climate science, critical zones, physical geographers, and a combined category for other science domains). A set of "builder" domains (computer/ cyberinfrastructure science, data management, software engineering, and a combined category for other expert domains) are also included. Three regression models are estimated, each with one category excluded (the coefficients on the other categories are interpreted relative to the excluded category). This allows for models that are in reference to an established field (geologist), an interdisciplinary domain (bio/ecosystem scientists), and a builders' domain (cyberinfrastructure experts).

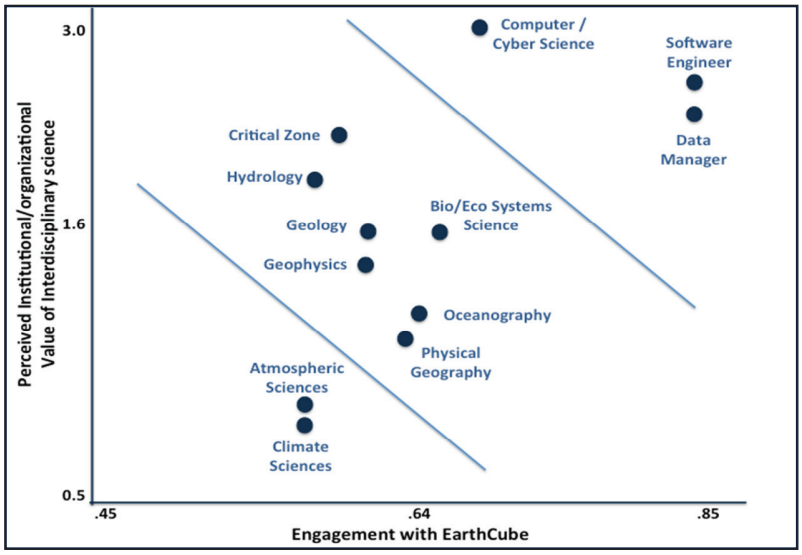

Figure 5: Support from Employer/Institution for Interdisciplinary Science and Engagement with EarthCube (N = 785 for the second version of the survey).

Note: Engagement with EarthCube by organizations and institutions that individuals are part of is correlated with support for interdisciplinary science. Cyber professionals, builders of infrastructure, populate the upper right zone above the contour line and are most positive.

Independent Variables	*Dependent Variable: Engagement with EarthCube[1] (B / Standard Error)*							
	Model 1 N = 1,508		*Model 2 N = 1,508*		*Model 3 N = 1,508*		*Model 4 N = 1,248*	
Constant	0.120	0.158	0.070	0.185	1.394	0.206***	0.467	0.277
Years of Experience	0.230	0.036***	0.230	0.036***	0.231	0.036***	0.266	0.041***
Female/Male (0,1)	–0.071	0.076	–0.071	0.076	–0.065	0.076	–0.028	0.085
International/ Domestic Org (0,1)	1.045	0.103***	1.044	0.103***	1.050	0.103***	1.070	0.124***
Atmospheric Science (0,1)	–0.597	0.123***	–0.547	0.157***	–1.883	0.176***	–1.621	0.198***
Geology (0,1)			0.051	0.138	–1.285	0.160***	–1.479	0.195***
Geophysics (0,1)	–0.235	0.126	–0.185	0.159	–1.521	0.178***	–1.057	0.175***
Hydrology (0,1)	0.215	0.166	0.266	0.192	–1.071	0.208***	–1.255	0.196***
Oceanography (0,1)	–0.470	0.124***	–0.420	0.156**	–1.756	0.176***	–0.873	0.225***
Bio/Eco Systems Science (0,1)	–0.050	0.139			–1.336	0.187***	–0.646	0.305*
Climate Science (0,1)	–0.938	0.167***	–0.887	0.192***	–2.224	0.209***	–1.964	0.232***
Critical Zone (0,1)	0.455	0.256	0.506	0.274	–0.830	0.286**	–1.170	0.205***
Physical Geography (0,1)	–0.263	0.253	–0.212	0.270	–1.548	0.282***	–1.409	0.319***
Other Scientists/ Others (0,1)	–0.162	0.145	–0.112	0.174	–1.448	0.192***	–1.254	0.213***
Computer/Cyber Scientists (0,1)	1.336	0.162***	1.387	0.188***				
Data Manager (0,1)	0.420	0.199*	0.470	0.22*	–0.865	0.235***	–0.698	0.262**
Software Engineer (0,1)	0.610	0.315	0.610	0.315	0.608	0.316	0.585	0.355
Other Experts (0,1)	0.167	0.174	0.218	0.199	–1.119	0.215***	–0.931	0.235***
Importance/ Ease Within Fields[2]							–0.101	0.174
Importance/ Ease Across Fields[2]							0.560	0.164***
EC Engagement Support Scale[3]							0.893	0.149***
F		23.268***		23.268***		23.040***		17.955***
Adjusted R Square		.19		.19		.19		.21

Table 2: Predicting Engagement with EarthCube.

Note: The engagement in EarthCube of cyber domains is significantly higher than the geo fields. Among geo and interdisciplinary domains, atmospheric science, climate science, and oceanography are less likely to be engaged. The interdisciplinary gap between importance and ease and support for engagement in EarthCube are positive predictors of actual engagement.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

[1] Six item ordinal scale for increasing degrees of engagement with EarthCube.

[2] Calculated gap between perceived "importance" and "ease".

[3] Two-item scale – Cronbach's alpha = .79.

Beginning with the control variables in **Table 2**, all three models show that more experienced researchers tend to be more engaged with EarthCube at this baseline. This is good news in that EarthCube is attracting experienced individuals at the outset who can provide important leadership. Over time, however, it will be a concern if this isn't followed by engagement of the next generation (another potential failure mode). Gender is not a factor with respect to engagement in EarthCube, which is a favorable finding. EarthCube is

a U.S. based initiative, and researchers affiliated with a U.S. organization or institution are more positively and significantly engaged with EarthCube than respondents outside the U.S. Over time, however, success will depend on broad international participation.

Turning to the fields, Model 1, has geology as the excluded variable. Geology was selected as a well-established field with particular challenges in the sharing and reuse of physical samples. Thus, the regression coefficients should be interpreted relative to geology. We don't specify a hypothesis since geologists could be expected to be more involved in EarthCube since the need is greater or less, since sharing and reuse is difficult. Among other traditional disciplines atmospheric/space scientists and oceanographers are significantly less likely to be engaged than geologists. Among the more interdisciplinary domains, climate scientists are less likely to be engaged. Interestingly, all three are areas of science that use sensors and modeling, which are challenging to share and reuse. Two of the three "builder" roles (computer/cyberinfrastructure scientists and data managers) are more likely to be involved. In Model 2, bio/ecosystem science is excluded as an illustrative interdisciplinary domain and the results are comparable even though they are in reference this time to bio/ecosystem science.

In Model 3, computer/cyberinfrastructure science is excluded as the lead builder domain. Here all roles other than software engineers are significantly less likely to be engaged in EarthCube than the computer/cyberinfrastructure scientists. Model 4 adds two gap variables (gap between importance and ease) from the first barrier (perceived need for change) and the support variable from the second barrier (support for engaging by colleagues and employing organization).[4] Holding constant the control variables and the various fields, we see that the interdisciplinary gap between importance and ease is a stronger driver of engagement than the gap within fields. Thus, sharing and reusing data across fields and disciplines is a stronger driver than sharing and reuse within a given field. We also see that internal support for engagement with EarthCube is a positive and significant predictor of engagement. Even though the levels of support (from the employing organizations and from colleagues) are relatively low, the workplaces where there is support do have a positive and significant impact on engagement. Thus, atmospheric/space weather science, oceanography, and climate science are less likely to be involved than other geoscience and interdisciplinary domains while builders of cyberinfrastructure are much more likely to be involved in EarthCube at the baseline – indicating that builders are initially more engaged than potential end-users.

8 Discussion

The gap between importance and ease in data sharing is a potential driver for engagement in EarthCube. The absence of such a gap would be a major barrier for any cyberinfrastructure initiative. This gap is consistent across geoscience fields, as well as among cyberinfrastructure experts, software developers, and data managers. Of course, end-users will only engage in EarthCube if they see it as an effective way to close that gap.

The relatively limited support from employing organizations and colleagues more broadly that is evident in the discussion of barrier two and in regression model 4, is a challenge for EarthCube. The results indicate that the issue is more pronounced with support from the employing organization than with colleagues. If the organizational and institutional infrastructure is not aligned in support of data sharing, then progress will be constrained. Although we did not ask respondents what obtaining such support would require, this was a focus of discussion in the various end-user workshops. Participants pointed to the need for systems to provide credit for reuse of data, similar to citation counts for publications, so that data sharing could be taken into account in tenure and promotion decisions for university faculty and career advancement for professional staff in data facilities. For developers of software, models, and other data products, parallel credit is also important and it will be important to better understand if there are such credit mechanisms in place where people do report strong organizational and institutional support. Participants also highlighted the need for funding, time and other forms of support for attaching metadata to aid in the discovery and reuse of data.

It is not surprising that there would be a gap between ease and importance in the early stages of this initiative. Yet addressing the gap will not be easy since institutional change involves changing institutional

[4] Although the two gap variables are highly correlated, which violates an assumption of OLS regression, additional tests were run including each separately and entering them in alternate order, with identical results. Only the EarthCube support variable is included since the interdisciplinary support variable involves about half of the cases, sacrificing statistical power and complicating comparisons across the models. The results with this variable indicate, however, that specific EarthCube support is a stronger driver than general interdisciplinary support.

leaders' and scientists' attitudes and behaviors towards data sharing. A small example of progress along these lines was evident in the end- user workshop for early career scientists. Sixty-eight doctoral students, postdocs, and assistant professors in their first few years participated in the workshop. (They were among the end-user workshop participants surveyed in the stakeholder survey). Afterwards, the participants were offered the chance to have a letter written to their department chair and dean indicating the strategic importance of the workshop and the merits of having been invited to attend. Most of the participants indicated an interest and many reported positive one-on-one conversations with institutional leaders afterwards. Much more is needed for institutions to be seen as valuing data sharing, but this example illustrates that the potential barrier is not insurmountable.

Finally, the evidence clearly indicates that the initial engagement in EarthCube is stronger for cyberinfrastructure experts, software developers, and data managers than it is for geoscientists. There can be carrots and sticks to motivate both the "build it" and the "they will come" aspects of open data initiatives. With sufficient carrots, builders will develop the cyberinfrastructure, which they have illustrated by their early participation in EarthCube. The National Science Foundation responded to the risk of builders getting too far out front by adding end-user workshops. More recently, the EarthCube governance structure has provided explicit avenues for the engagement of scientists – particularly around the identification of "use cases" to guide the development of the cyberinfrastructure. Observers at EarthCube governance sessions, including members of our research team, note that there quickly surfaced widely different definitions of what constitutes a use case. There is emerging recognition that there is value in diverse types of use cases and scholarly recognition for a more systematic and modular approach to use cases (Jacobson, Spence, Kerr, 2016). Beyond seeking greater stakeholder engagement, it is important to advance the enabling mechanisms, such as the utilization of use cases. In further research and future rounds of data collection, it will be important to track the degree to which the engagement of the end-users is expanding and generating actual data sharing behaviors.

9 Conclusion

This analysis of the EarthCube initiative in its early stages reveals there is a substantial gap between the perceived importance of data sharing in the geosciences and the perceived ease of doing so. Rather than being a barrier, this is a potential driver for EarthCube – particularly since the gap is consistent across a wide range of fields and domains. However, while assessment of builders' and end-users' needs is necessary, it is not sufficient.

Also important is organizational and institutional support for end-users to engage in the initiative. The baseline data presented here suggest that end-users view organizational and institutional support as moderate. This is important to address because low support may reduce scientists' interest in participating. While organizational and institutional change doesn't happen easily, each step in the direction of lowering barriers not only removes a barrier but produces an additional driver. Given the requirements of the 2013 OSTP data sharing memo and other forces driving increased data sharing, these organizational and institutional considerations pose a crucial set of challenges for data science.

Finally, the data show that in this early stage, builders are more engaged than end-users. This represents an important risk to EarthCube's success. Builders are essential, but the initiative needs comparable engagement and even leadership from the end-user scientists. Thus, while it seems likely that the data and cyberinfrastructure experts will "build it," the real concern is that geoscientists "will not come."

All three potential barriers – perceived need, perceived support, and distribution of engagement – represent key baseline indicators that inform ongoing efforts to avoid a key failure mode in cyberinfrastructure initiatives. The NSF is not blindly building EarthCube, hoping that the geoscientist will come and share data, tools, and models – there is clear evidence of a perceived need. The challenge now is to progressively lower the barriers and increase engagement in the building process, so that as it is built and as it evolves, they will come.

Supplementary Files

The supplementary material for this article can be found as follows:

- **Supplementary File 1:** Appendix. http://dx.doi.org/10.5334/dsj-2016-008.s1

Acknowledgements

Support from the National Science Foundation is deeply appreciated. Support for this research has been provided through NSF OCI RAPID 1229928, "Stakeholder Alignment for EarthCube," and NSF GEO-SciSIP-STS-OCI-INSPIRE 1249607, "Enabling Transformation in the Social Sciences, Geosciences, and Cyberinfrastructure."

Competing Interests

Nine of the twenty-four co-authors of this article have served as facilitators or active contributors to the formation of EarthCube. When a co-author was serving as a facilitator, one or more other members of the research team attended the same event as observers. Established methods of action research and participant observation informed the work of all co-authors with respect to the EarthCube initiative.

References

Deming, W E 1986 Out of the Crisis, Massachusetts Institute of Technology. *Center for advanced engineering study, Cambridge, MA* (510).

Finholt, T A and **Birnholtz, J P** 2006 If We Build It, Will They Come? The Cultural Challenges of Cyberinfrastructure Development. In: *Managing Nano-Bio-Info-Cogno Innovations*. Springer, pp. 89–101. DOI: http://dx.doi.org/10.1007/1-4020-4107-1_7

GEO Vision Report 2009 Washington, D.C.: National Science Foundation. Retrieved from: http://www.nsf.gov/geo/acgeo/geovision/nsf_ac-geo_vision_10_2009.pdf.

Hendrix, P E 1999 Build It, and They Will Come. *Marketing Management, 8*(4): 31.

Jackson, S J and **Buyuktur, A** 2014 Who Killed Waters? Mess, Method, and Forensic Explanation in the Making and Unmaking of Large-Scale Science Networks. *Science, Technology & Human Values*. DOI: http://dx.doi.org/10.1177/0162243913516013

Jacobson, I, Spence, I and **Kerr, B** 2016 Use-Case 2.0: The Hub of Software Development, *ACM Queue, 14*(1): 94–13. DOI: http://dx.doi.org/10.1145/2890778

Katz, M L and **Shapiro, C** 1994 Systems Competition and Network Effects. *The journal of economic perspectives,* 8(2): pp. 93–115. DOI: http://dx.doi.org/10.1257/jep.8.2.93

Kinsella, W P and **Parker, T** 1982 *Shoeless Joe*. Houghton Mifflin New York.

Lawrence, B S 2006 Organizational Reference Groups: A Missing Perspective on Social Context. *Organization Science, 17*(1): 80–100. DOI: http://dx.doi.org/10.1287/orsc.1050.0173

Lewin, K 1939 Field Theory and Experiment in Social Psychology: Concepts and Methods. *American journal of sociology*: 868–896. DOI: http://dx.doi.org/10.1086/218177

Markus, M L and **Keil, M** 1994 If We Build It, They Will Come: Designing Information Systems That People Want to Use. *Sloan Management Review, 35*(4): 11.

Mayernik, M S, Wallis, J C and **Borgman, C L** 2013 Unearthing the Infrastructure: Humans and Sensors in Field-Based Scientific Research. *Computer Supported Cooperative Work (CSCW), 22*(1): 65–101. DOI: http://dx.doi.org/10.1007/s10606-012-9178-y

OSTP Public Access Memo 2013 (February 22) Retrieved from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Ribes, D and **Finholt, T A** 2008 Representing Community: Knowing Users in the Face of Changing Constituencies. In: *Proceedings of the 2008 ACM Conference on ComputerSupported Cooperative Work (CSCW)*. New York: ACM, pp. 107–116. DOI: http://dx.doi.org/10.1145/1460563.1460581

Stakeholder Alignment Collaboration: Cutcher-Gershenfeld, J, Baker, K S, Berente, N, Carter, D, DeChurch, L, Flint, C, Gant, J, Gershenfeld, G, Grant, B, Haberman, M, King, J L, Knight, E, Lawrence, B, Lewis, S, Lopez, P, Mayernik, M, Mcelroy, C, Mittleman, B, Nichol, V, Nolan, J M, Pak, S, Ruengvisesh, D, Shin, N, Thompson, C A, Winter, S and **Zaslavsky, I** 2015 (working paper) Internal Alignment for Multi-Stakeholder Consortia.

Tintinalli, J E 2008 Build It and They Will Come. *Emergency Medicine Australasia, 20*(3): 193–195. DOI: http://dx.doi.org/10.1111/j.1742-6723.2008.01090.x

Waitz, I A, Townsend, J, Cutcher-Gershenfeld, J, Greitzer, E and **Kerrebrock, J** 2003 Aviation & the Environment. *Report to the United States Congress*.

Yuan, T, Crowley, J, Asunka, S, Chae, H S and **Natriello, G** 2010 Build It and They Will Come? *International Journal of Advanced Corporate Learning (iJAC), 3*(3): 39–44.

How to cite this article: Cutcher-Gershenfeld, J, Baker, K S, Berente, N, Carter, D R, DeChurch, L A, Flint, C C, Gershenfeld, G, Haberman, M, King, J L, Kirkpatrick, C, Knight, E, Lawrence, B, Lewis, S, Lenhardt, W C, Lopez, P, Mayernik, M S, McElroy, C, Mittleman, B, Nichol, V, Nolan, M, Shin, N, Thompson, C A, Winter, S and Zaslavsky, I 2016 Build It, But Will They Come? A Geoscience Cyberinfrastructure Baseline Analysis. *Data Science Journal,* 15: 8, pp. 1–14, DOI: http://dx.doi.org/10.5334/dsj-2016-008

Submitted: 21 February 2016 **Accepted:** 02 June 2016 **Published:** 11 July 2016

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

OPEN ACCESS