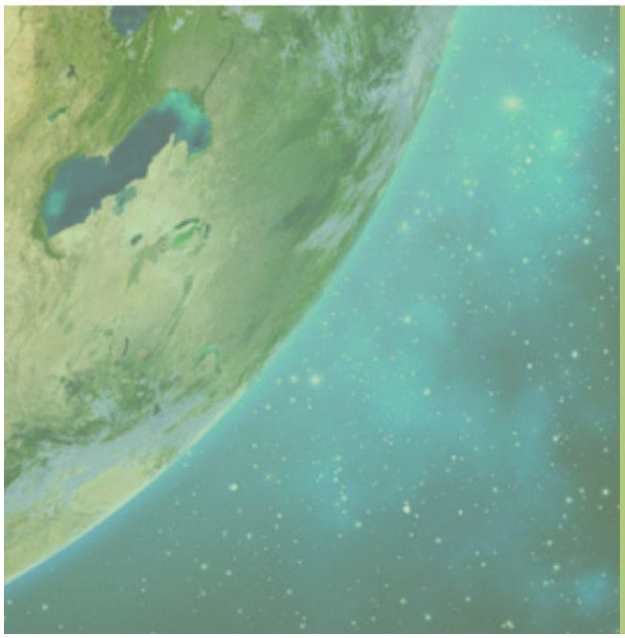
PROJECT REPORT

Sustainability models for integrated digital Earth Science

Submitted to: EarthCube

May 31, 2022



EarthCube
Transforming Geosciences Research

PREPARED BY
Arika Virapongse*
James Gallagher*
Basil Tikoff*
Peter Cornillon
Brooks Hanson
Rebecca Koskela
Susan Shingledecker
Chad Trabant

CONTACT*
av@middlepatheco.com
jgallagher@opendap.org
basil@geology.wisc.edu

Sustainability models for integrated digital Earth Science

A project report for EarthCube

Executive Summary

The NSF Earthcube program sponsored a study of Earth Science infrastructure project sustainability. This report documents that study and its results. Eleven data infrastructure projects, most focused solely on the Earth Sciences, were studied by conducting interviews with a person who was or is currently a leader in the project. The study reveals that, even though the projects studied varied significantly in many ways, sustainable projects share common traits, approaches, and challenges.

Lessons learned

- There are two parts to sustainability: 1. sustainability of the project; and 2. sustainability of the product/mission.
- We noted the existence of three different types of digital data products: Database, Framework, and Middleware. Identifying a typology was helpful for determining some practices and pitfalls for sustainability.
- The transition from the initiation phase to operations phase of a digital infrastructure project is often a major stumbling block. This occurs because the attributes (skills, business models, staff, etc.) needed to be successful in the initiation phase are different from those needed at the operations phase.
- None of the studied projects began with a formal governance model. Instead, each project adopted a governance model over time.
- The three types of data infrastructure projects require different levels of interactions with their communities and approaches to involve the community in governance. Database projects in particular require engagement by trusted disciplinary scientists. Framework projects were inseparable from their community. Middleware projects worked on building community trust to a lesser degree.
- There is likely no “one size fits all” approach to sustainability within the Geosciences. The context and goals of each project determine the different types of approaches and business models needed to increase the potential for sustainability.

Recommendations

- US federal funding agencies should provide decade-long development support for projects to successfully progress through the initiation stage, particularly for science domain communities that do not yet have digital infrastructure.
- US federal funding agencies should create funding programs that more directly meet the needs of science data infrastructure. Data infrastructure projects are mission-driven, often addressing broader complex system issues. Projects must work on short-term time scales

because of the nature of current research funding models, but all have the long-term burden of sustaining and supporting open science, interoperability, and community building. These additional long-term activities are often unfunded, because they do not create the products required by short-term funding.

- NSF should consider implementing a program similar to I-CORPS, but targeted at projects that aim to become (actually or effectively) not-for-profit organizations. Currently, there is little assistance to help projects develop as not-for-profit organizations. This could help bridge the critical gap between project inception and the period when the initial leaders must first develop a formal governance mechanism.

- An NSF RCN-like project should be funded to bring together Database and Framework projects to discuss and innovate solutions for appropriate, long-term business models. These two project types require unique solutions because they inherently have a commitment to and contribution from a disciplinary community. Further, they are essential to current scientific practice. Thus, finding an appropriate model for a sustainable future for these types of projects would be highly beneficial.

- Projects should build community engagement into their initiation phase that is appropriate to the type of project. Projects should also be cognizant of the need to develop formal governance structures and a business model toward the end of an initiation phase.

- The Geoscience community needs to continue community-scale interactions (conversations, workshops, meetings, etc.) - in the absence of EarthCube - to pursue a sustainable future for data infrastructure.

Authors/contributors

Arika Virapongse, James Gallagher, Basil Tikoff
Peter Cornillon, Rebecca Koskela, Susan Shingledecker, Chad Trabant, Brooks Hanson

Acknowledgements

Funding was provided by the National Science Foundation EarthCube Council of Funded Projects for a project entitled, "Sustainability models for integrated digital Earth Science beyond EarthCube."

1. Introduction

This report describes the results of a study on the sustainability of Earth Science data infrastructure projects. The study aimed to address core issues of 'project sustainability' by examining projects relevant to EarthCube that have successfully outlasted their initial period of inception. The original motivation for the work were discussions in the EarthCube Council of Funded Projects.

We explored the question: ***"What are the challenges, lessons learned, and recommendations for sustainability of Earth Science data initiatives?"*** By examining efforts that grappled with organizational sustainability, we intended to find common traits shared across a spectrum of projects that address issues in the Earth Sciences. Such insights can be useful for any Earth Science data infrastructure initiative seeking to develop their sustainability, and for funding agencies with goals to better support Earth Science.

Intended outcomes of the study were/are:

- An internal EarthCube report on different models for sustainability
- A paper in a peer-reviewed, international journal highlighting findings and recommendations for projects that intend to be sustainable.

This report completes the first of these bullet items and includes some preliminary discussion points. A subsequent peer-reviewed publication will include more detailed recommendations and conclusions. We have broken the study into these two distinct phases to provide the subject interviewees with an opportunity to comment on the information in the Results section before deriving recommendations from that material.

1.1 Background

In this report, we use the term 'project' to describe a group of people combined into or acting as one body—otherwise known as a 'corporation'[1] or organization that often includes various legal or governmental rights. Such a body is created so that it can be separated from the individuals that use and work within the body. This idea is fundamental to how we define sustainability here: the continued existence and goals of a project that continue despite the change in individuals.

Importantly, sustainability is about a project staying relevant beyond the initial seed or start up funding. For a project to continue over time, the project must be able to demonstrate its value to an agency, its stakeholders, and/or clients. Can sustainability after initial funding ends be used as a metric to assess the value of funded projects?

In this study, we define community as stakeholders, end users, collaborators, and anyone participating in a project outside of paid staff. Science is arguably a community-based enterprise at its core: Scientific progress often occurs through collaborative groups, peer review, meetings, workshops, and discussions. Likewise, Earth Science data and infrastructures initiatives have often used community-based approaches as either part of their strategy or as the main basis for what they do. While there are studies of community-based initiatives overall,[2] there is little known about how well these community-based approaches work for Earth Science data and infrastructures initiatives, or how this approach might contribute to the sustainability, growth, and relevancy of these initiatives.

In this study, we examined 11 data infrastructure projects, primarily in Earth Science but also including one broader effort, to evaluate the qualities, challenges, vision, and strategies that affected their overall sustainability.

2. Methods

This project was designed as a social science team research project. Interviews with 11 key informants were used as the source of primary data. The project consisted of two phases with the first lasting ten months (February-November 2021) and the second lasting five months (January to May 2022); these phases were determined through oversight by the EarthCube office (the research funder). Each interview was conducted by two or three core team members. The recording, transcript, and notes were made available to the full analytical team (8-10 individuals total), who annotated the transcript (known as "coding" in qualitative analysis). In a subsequent meeting, the full team discussed their reflections about the interview. The full procedure outlining the methodology of these interviews is provided in Appendix A1.

[1] The American Heritage® Dictionary of the English Language, 5th Edition.
[2] Ceptureanu, Sebastian I., Eduard G. Ceptureanu, Cristian E. Luchian, and Iuliana Luchian. 2018. "Community Based Programs Sustainability. A Multidimensional Analysis of Sustainability Factors" *Sustainability* 10, no. 3: 870. https://doi.org/10.3390/su10030870

3. Results

The results are organized by 5 main topic areas: Leadership and Decision-making, Community engagement, Benefits, Business model, and Sustainability. The sub-topic areas were elicited based on the clustering analysis approach.

The text in Bold indicates the main point of a cluster of quotes (representing raw data). In parentheses, the frequency of interviewees included in the cluster is noted with "n", and the specific interviewees in the cluster are noted using the convention of Interviewee Number-Subsample group (D=Database, F=Framework, M=Middleware). Next, a summary to describe the cluster is included. Finally, a quote from one of the interviewees in the cluster may be included.

3.1 Leadership and Decision-making

3.1.1 What does Governance look like?

3.1.1.1 Governance must be intentionally created (n=4; I2-F, I4-D, I6-F, I10-D). Governance as an intentional process leads to more sustainability. Interviewees groups noted the importance of putting resources into developing and establishing an effective governance structure, particularly to integrate community input into the project.

Quote from I6: "*setting up the governance was just really important if you want the organization to be sustained. Yes, there's the money but you also need to have the organizational structure that can sustain it*"

3.1.1.2 An established process for making decisions is needed (n=4; I2-F, I4-D, I5-F, I6-F). Voting was used by the Framework sub-sample groups. A consensus approach was used by the Database project, which was notably a loose collection of researchers working together on the project.

3.1.2 Balancing between top-down and bottom-up governance

Top-down often consists of project staff and selected decision-makers. Bottom-up often represents stakeholders in a project, end users, and general community that interacts with, contributes to, and has interests in the project.

3.1.2.1 Top-down and bottom-up governance must be balanced in different proportions (n=6; I5-F, I6-F, I7-D, I8-D, I9-D, I11-M). The interviewees used qualifiers like "balance" and "mixture" to describe their governance model. Interviewees referred to the role of top-down decision-making for handling the funding, funders, requirements of the funders, bottom-line of the project, and supporting the community. In contrast, bottom-up governance had the

role of providing ideas, innovating along the edges of the project, and implementing/ experimenting with products.

Quote from I7: "*It was an extremely healthy mixture of both of those models that—clearly at the corporate level—I and program managers were responsible for, you know, a lot of the putting forward of ideas developing initial budgets. But it was then a process of going through and making sure that those withstood the review of the community. And it wasn't until they had the approval of the community that they went forward*"

3.1.2.1 Leaning more bottom-up is better for some projects (n=4; I2-F, I3-M, I6-F, I7-D). These projects were all heavily community-oriented. They recognized the importance of community ownership, the role of community in their project relevancy (maintaining funding streams), and for meeting their mission as a project.

3.1.2.1 Leaning more top-down is better for some projects (n=3; I3-M, I5-F, I10-D). Top-down governance provides necessary oversight and good decision making (with carefully selected leaders) for a project. It also helps the community with things that they don't do well, like complete tasks that require long-term commitment and considerable investment. One interviewee emphasized that democracy can be overdone.

3.1.3 Who makes decisions?

Governance is often composed of individual leaders like hired staff and community champions, and small groups like Boards and working groups.

3.1.3.1 Good leadership characteristics can vary according to the needs, stage, and goals of a project (n=8; I2-F, I3-M, I4-D, I6-F, I7-D, I8-D, I9-D, I11-M). Some leaders are chosen to provide credibility and star power, so they can more easily capture funding; however, these characteristics can be at odds with community ownership. Domain expertise/experience and being competent at doing hands-on work are also valued. Leadership needs can vary depending on what the project is aiming to accomplish for their mission and/or at a particular stage.

Quote from I6: " *if that one head is very powerful and can raise money like this, then there's real trade offs there, right, because it's, it's like well you get a powerhouse who can just write checks or do you try to make it into something which is [community] owned*"

3.1.3.2 Having mechanisms to put the right people into leadership roles (n=7; I1-M, I2-F, I3-M, I4-D, I6-F, I9-D, I10-D). Succession plans are key to success for a project, especially in case leadership is suddenly lost. For paid positions, new leaders are found/created by recruiting through personal networks, promoting someone already doing the work, and rallying around

leadership changes. For volunteer positions, interviewees mentioned having formal structures that rotate people on and off, and current leaders nominating/recruiting new leaders.

3.1.3.3 Diversity in leadership and decision-making is needed (n=4; I4-D, I6-F, I7-D, I10-D). Interviewees spoke to the need and desire to represent the professional background, domain expertise, and skill sets of the users of their database projects in their decision-making. The Framework project spoke about conflict that occurred when introducing more career-level, nationality, and cultural diversity into leadership.

3.1.3.4 Creating formal groupings helps to bring community perspectives into decision making (n=4; I3-M, I4-D, I5-F, I8-D). Formal groups included: policy committee, advisory board, executive committee, working groups, and strategic members.

3.1.3.5 There are gaps in leadership skills and training for running these projects (n=2; I3-M, I8-D). Leadership skills were often learned on the job, since leaders are often selected for their domain expertise and not for leadership skills. Projects within academia/science require a unique leadership model that is not well understood.

3.2 Community Engagement

3.2.1 Community participation strategies

(n=8; I1-M, I2-F, I3-M, I4-D, I5-F, I6-F, I8-D, I9-D). Most strategies consisted of identifying the benefits that people could get from participating (e.g., learning, access to new resources, having power over design of a product), as incentives for participation. Giving people funding (and making it a requirement to participate in community activities) was also used to seed participation.

3.2.2 How a community benefits the project

(n=7; I1-M, I3-M, I4-D, I5-F, I6-F, I7-D, I11-M). The community helped projects stay positioned. People in the user community developed products that projects could point to as successes. They also created new products to link a project's product to other products–broadening its benefits and use.

Quote from I5: "*because their missions are big you need as much credibility and as much lowering the risk as possible and that's why they go back to: let's work with the community of experts so I make sure that my positioning is accurate*"

3.2.3 Community outreach and engagement strategies

(n=6; I2-F, I3-M, I4-D, I9-D, I10-D, I11-M). Outreach consisted of targeted recruitment, domain conference attendance (including having a booth), spending time with researchers where they do their field research, and hosting "user conferences". Time and effort was spent specifically to interact with the user community.

Quote from I9: *"We can poll our researcher community really easily. [Founder and previous lead] was adept at doing that, so whether it was meetings of opportunity, and making sure she went and shook hands of all the people that submitted data, or whether she was actually on the deck of a ship with them, you know that was a really great advantage that she had...most of the staff have been out to sea throughout their careers."*

3.2.4 Community engagement is challenging

(n=3; I8-D, I9-D, I11-M). Members of user communities are busy people, so it's hard to get time with them. A project must devote intention and resources (e.g., a support person) to cultivate and maintain connections with end users.

3.3 Benefits

3.3.1 Measuring value, benefits, and quality of the project is important

(n=10; I1-M, I2-F, I4-D, I6-F, I7-D, I8-D, I9-D, I10-D, I11-M). Almost all interviewees in our sample group mentioned ways that benefits were measured in their project (e.g., number of products, number of participants). The value of open source is harder to measure, because it is difficult to track who is using tools. For community-centric efforts, tangible outputs are important. One way to measure qualitative benefit is to look and see if the target group(s) can achieve their objectives.

Quote from I1: "*we don't really know how many of the servers are out there, we have looked into different ways of trying to gather this information. And in general, there's pretty strong pushback from the data provider community—that collecting that sort of information would be a privacy violation*"

3.3.2 Value Proposition

The different ways that projects envisioned the value that they provided.

3.3.2.1 Creating infrastructure that supports interoperability (n=9; I1-M, I2-F, I3-M, I4-D, I5-F, I6-F, I8-D, I10-D I11-M). Most projects in the sample group aimed to support and intersect more than one discipline and sector–through activities like meetings and/or developing new platforms, middleware, and databases. Meetings (and associated activities) are often a key activity, because they bring people of different domains/sectors together.

Quote from I3: "*I think there's real value in trying to look for infrastructure that can support more than one discipline.*"

3.3.2.2 Filling gaps and exploring new frontiers (n=4; I2-F, I3-M, I6-F, I8-D). Projects found opportunities to create a place for a new community, take advantage of new technology (like

the internet), fulfill domain specific needs (create databases), and fill in a gap in science workflows.

3.3.2.3 Usefulness and problem solving (n=4; I2-F, I3-M, I5-F, I8-D). Projects often focused on solving specific problems, and meeting an essential need. This was achieved through bringing together people (communities) to work on problem solving, as well as through tool development.

3.3.2.4 ***Capacity building (n=4; I3-M, I4-D, I7-D, I8-D).*** Education, training, and learning opportunities for the community were often key elements of projects because they encouraged people to participate in the community and also get more value out of new technology and tools.

3.3.3 Enhancing benefits

(n=8; I3-M, I5-F, I6-F, I7-D, I8-D, I9-D, I10-D, I11-M). Projects often looked for ways to broaden and/or accelerate benefits. Some ways that this was done included: bringing new combinations of different stakeholders together (e.g., publishers and researchers), pushing out existing ideas to new domains, enabling the community, diversifying and expanding the group of users, and crowdsourcing to improve product quality.

Quote from I11: "*We make sure that it runs everywhere...it has to run on everything from an ARM processor to the biggest supercomputer in the world*."

3.4 Business model

3.4.1 Staff description

3.4.1.1 Expectations of staff (n=6; I3-M, I4-D, I6-F, I8-D, I9-D, I10-D). Staff must often work in more than one area and across domains; some were conduits to domain-specific end users. Technology efforts often sought individuals with improved software engineering skills, and sometimes for only part-time. Many projects depended on leaders and volunteers who had other hard-money jobs.

3.4.1.2 Turnover of staff was a challenge (n=4; I1-M, I4-D, I8-D, I11-M). Most projects work in a highly technical and specialized domain, and have a small staff. Losing staff often meant that a project lost an entire area of expertise, utility, and/or access to user groups. In addition, working on short project cycles, with academics, and with limited financial reserves often resulted in short-term staff.

Quote from I8: "*We've always hired really interesting people, and really interesting people only stay as long as the work is really interesting*."

3.4.2 Partnerships with other entities and stakeholders

(n=9; I1-M, I3-M, I4-D, I5-F, I7-D, I8-D, I9-D, I10-D, I11-M). Almost all of the interviewees in the sample group noted the importance of partnerships. Partnerships included: a consortium model (different projects or individuals working together towards a common goal), working across sectors (e.g., a nonprofit collaborating with universities), collaborating with different projects to leverage each other's strengths, and pooling together to access limited resources (e.g., grants). Some partnerships were "forced marriages", where a funder asked two related projects to join together.

3.4.3 Funding the project

3.4.3.1 Research grants were a primary source of funding (n=11; I1-M, I2-F, I3-M, I4-D, I5-F, I6-F, I7-D, I8-D, I9-D, I10-D, I11-M). Research grants came from U.S. federal agencies, Industry, and Foundations. Longer grant periods were generally better for project development, since many projects were working on efforts that would likely take 10 years to mature. Good connection and alignment with the funder was often key to long-term funding success–some projects strategically cultivated this relationship. Projects often sought to sequence short term projects so that they served a larger longer-term goal. One weakness with the grant-driven model is that it's very hard to get grants; it is dependent on having someone in the project who is successful at writing funded proposals.

Quote from I6: "*NIH will keep funding you, and I said, oh yeah I know they will. But I don't want them to, right? I actually want to think about this because...sustainability of these digital infrastructures is a real issue*"

3.4.3.2 Funding through donations and membership subscription models were not often successful (n=5; I1-M, I4-D, I5-F, I6-F, I7-D). Interviewees recognized that donation and membership subscriptions were potential sources of income. For most projects, individual-level donations and membership subscriptions are unlikely to work. One project has been able to use an institutional-level membership model successfully as part of their funding stream, while another tried this model and abandoned it early on.

3.4.3.3 Contracting was a complement to grant funding (n=4; I1-M, I7-D, I8-D, I11-M). Contracting was often conducted with the US federal government and agencies (never NSF), such as the Department of Defense and NASA, as an important counterpart to funding from research grants.

3.4.3.4 Selling services and products *(n=4; I1-M, I5-F, I8-D, I11-M).* Middleware projects can effectively sell a product(s). Some projects did try to sell access to a tool, but it was not as successful. Other types of projects can sell services.

3.4.3.5 Funding from Foundations and Private Industry is difficult to attain *(n=4; I4-D, I7-D, I8-D, I10-D).* Private funding was not a good source of funding for Database projects, although it was considered and attempted. While a few of these projects were able to leverage foundation funding, they were the exception.

Quote from I4: "*I've tried and I cannot find any [foundations] that seem aligned with what we want to do.*"

3.4.3.6 Mixed models for funding were often used *(n=4; I6-F, I7-D, I8-D, I9-D).* Many projects used mixed models for funding. They were aware that each source of funding had its limitations, so they leveraged different sources according to their strengths. A project might start out with grants as their main source of funding, but then they often evolved into a mix of grants and contracts, for example.

3.4.3.7 Funding from Meetings (n=2; I2-F, I6-F). Meetings can generate income through registration fees and sponsorship. But such a mechanism requires the capacity to organize such large events, as well as infrastructure to keep interest/work/cohesion going between the meetings (e.g., working groups).

3.4.4 In-kind support

(n=2; I4-D, I5-F). Projects were often initiated by university faculty who had full time employment. Projects also often depended on volunteers with other employment.

3.4.5 Large host organizations helped to cover administrative costs

(n=2; I3-M, I8-D). Administrative costs can be disportionately high for these relatively small projects. Working under the umbrella of a large organization like universities helped to provide needed business infrastructure, such as help with taxes, payroll, and legal support.

3.4.6 The academic business world is unique

(n=2; I8-D, I9-D). Projects found that conventional business approaches did not work for their specific needs. And likewise, it was important to understand academic culture to do well as a business.

Quote from I8: "*it's not just the money...it's the management structures and they are a little unique to academia...you cannot port a business model—a business structure—into academia, with great ease*"

3.5 Sustainability

3.5.1 Tension between goals/sectors/domain areas

3.5.1.1 Community/Research versus tool development/facilities (n=6; I1-M, I3-M, I6-F, I7-D, I8-D, I11-M). Projects noted that they had to make an intentional decision between devoting their limited resources to research or community-oriented goals/activities that directly benefited a broad range of individuals and the development of a tool, platform, or technical infrastructure that might have limited benefits for the community members involved in the

project. One interviewee from a Framework project noted that building tools could create competition between members in their community. Technology is expensive to develop and sustain; resources can go a much longer way when devoted to community-oriented infrastructure.

Quote from I1: "*it felt like we were undercutting the scientific work to keep the technology funded. And we weren't even able necessarily to push the technology forward*"

3.5.1.2 Tension in leadership can lead to crises and failure (n=4; I2-F, I4-D, I10-D, I11-M). Poor leadership is when people are looking out for their own personal interests such as putting resources into passion projects or supporting their allies, and not putting the mission and health of the project first. There was often contention at the start of a project, and overcoming such leadership crises was key to the project's survival.

3.5.2 Forks in the road and reacting to major changes

(n=6; I2-F, I3-M, I5-F, I6-F, I8-D, I9-D) Many of the projects experienced failures, challenges, and not-great success. Projects with a heavy technology component must navigate societal--level changes in technology. Some projects experienced discord at their inception and yet went on to be successful. One project experienced an unsuccessful 'forced marriage'. Recognizing and confronting broader social issues (e.g., Diversity & Inclusion) can lead to major change in a project.

3.5.3 How to overcome and be resilient in "lean" times

(n=6, I1-M, I4-D, I6-F, I8-D, I10-D, I11-M) The lack of financial reserves was a major weakness for most projects because it impacted their ability to retain staff and withstand major shocks. Creating financial reserves with grant funding is not possible. Projects engaged in contracts to overcome this limitation, as well as leaned on a larger umbrella organization to relieve their management burden.

Quote from I8: "*I don't remember a year where we had enough money to pay the salaries at the end of the year. At the beginning of the year in the first five years, so you know, that was a little nerve wracking—that's still true*"

3.5.4 Early pioneers and innovators

(n=4; I3-M, I4-D, I5-F, I7-D, I8-D) Projects were often embarking on activities within a context that had never been attempted before. While this gave the projects the advantage of being on the leading edge, it was also risky. Most projects were not well set up to withstand much risk (e.g., little financial reserves and subsisting on short grant cycles). Being open to unusual new things, like partnering with the military or publishing companies, creating ways to build financial reserves, or allowing their community to take the lead, often provided the advantage of allowing them to capture new opportunities.

Quote by I3: "*to the extent that I know anything about computer science, it's all seat of the pants*"

3.5.5 Relevancy

3.5.5.1 Successfully staying relevant and essential (n=6; I3-M, I4-D, I5-F, I7-D, I9-D, I11-M). Projects sought to maintain their relevancy over time, and their strategies often included staying well connected to their end user community and funders. When such stakeholders insisted that the project was essential for their work, that was a good indicator of sustainability.

Quote by I3: "*it was almost like a utility that you plug it into the wall, and people really just expect it to be there*"

3.5.5.2 Challenges of staying relevant and useful (n=7; I1-M, I3-M, I4-D, I6-F, I9-D, I10-D, I11-M). In the effort to stay relevant, projects had to balance their urge to meet everyone's needs, while maintaining their focus and mission. Likewise, projects were challenged with balancing their business bottom line with innovation, which was a necessary part of staying relevant. Middleware projects were further challenged with the requirement of keeping their software up-to-date, because technology is always changing. Most of the projects had to also prioritize the ability of their community to participate in the project.

Quote by I10-D**: "***You can't build a system in 1996, and expect that it still works well in 2020. Right? So you have to make these updates that are huge efforts. They don't work in incremental ways... that has been a big challenge to explain to the community and to NSF. That once in a while you need this kind of boost in funding to get something new...With incremental improvements, you run into the danger of becoming irrelevant or disconnected from the overall infrastructure development***."**

3.5.6 Stay focused

(n=5; I3-M, I4-D, I5-F, I7-D, I11-M). It was enticing for projects to veer away from their mission and focus, but with limited resources, that was detrimental to the project. All of these projects were highly tied to their mission, and it was a main reason for their existence.

Quote from I11: "*the mission is what it's all about...the reason I had to come back was that our mission was suffering*."

3.5.7 Defining, imagining, and planning for sustainability

(n=8; I1-M, I2-F, I3-M, I4-D, I5-F, I7-D, I8-D, I9-D). Most of the projects had a vision for sustainability and had put considerable thought into what it means. Sometimes funders could help steward projects towards sustainability. Starting with the intention of sustainability can shape an organization in important ways. One Framework project explicitly decided to not develop technology, and to instead develop a community-based process for developing technical specifications. Other projects

have taken steps to ensure that if the project ended, the data were safe. Most projects aimed towards successful transitions of leadership as part of their sustainability plan.

Quote by I2: "*I wanted there to be an organization [Project] without me leading it.*"

3.5.8 The landscape of research grant funding has changed

(n=4; I2-F, I3-M, I7-D, I10-D) Projects are focused on issues that are essential for science today, but research grants were not originally designed to support them. Funders take risks to help initiate these projects, but sustaining them through this same source of funding is not within their scope.

3.5.9 Keep an abundance mindset

(n=3; I1-M, I3-M, I6-F) Despite the limited funding resources available for projects, projects still needed to maintain a generous collaborative mindset because science-based projects rely on partnerships. Open source software development, in particular, is pushing forward a culture of sharing and must live up to their message—despite essentially creating their own competition.

3.5.10 Good documentation is important for organizations

(n=3; I1-M, I5-F,* I8-D*) Interviewees noted that documentation on how to make decisions and what decisions were made were important components of good governance. Similarly, documentation was noted as being important for database development.

3.5.11 Is it worth sustaining?

(n=2; I7-D, I8-D) Interviewees were introspective about the parts of their projects that should be sustained.

Quote by I7: "*a key component of sustainability is the quality of what you want to sustain...has to be something that's worth sustaining*"

4. Discussion

The results describe the study projects in two broad ways: 1. the kinds of challenges they had to overcome or manage; and 2. the unique aspects of their problem space that could be used as assets to increase stability.

4.1 Challenges

4.1.1 Funding

Research grant funding sources are not well suited for these projects. Because these projects are intended to be run long-term they do not fit the commonly used research model where a study is performed for a finite period (the true definition of a 'project') capped with the publication of results and recommendations for further work that will take place as part of a subsequent project. On the other hand, a sustained data infrastructure project, like a corporation or a research program, is intended to exist until it is no longer useful, with no fixed end date.

All of the projects examined in this study started with a grant from an agency that primarily funds scientific research projects. Most of the study projects engaged in mixed funding models in order to leverage the strengths of one to offset the limitations of another. How the projects managed this vary by project type. Middleware projects generally mixed contract work from both government agencies and private companies. One of three Middleware projects is, however, almost solely funded by the U.S. National Science Foundation (NSF). Database projects tended to derive the bulk of their funding from NSF. Framework projects use a mix of grants spread across agencies along with private funding sources, membership fees, and meeting fees.

4.1.2 Business Models

The projects studied are structured around (at least some of) service, interoperability, collaboration, academia, and community-based approaches, and these characteristics can make a project a poor fit for conventional for-profit business models. Thus, when projects tried to apply models used successfully by for-profit businesses, it was not enough for success. In one case, there was substantial financial loss and organizational instability. Techniques like fee-for-service and membership fees are difficult for these projects to manage; they do not often lead to sufficient funds to run the organization alone. Indeed, the culture of academia and science, which is highly collaborative, is not often receptive to *quid-pro-quo* approaches. These revenue generating mechanisms, however, can augment other sources of funding, and the combination can yield sufficient funds for stability. In addition, some research suggests that the sustainability models for not-for-profit corporations may be applicable.[3]

4.1.3 Governance

The three categories of projects approached governance in very distinct ways. The Middleware projects used the least amount of outside community-driven input for their governance. Instead, they all relied on a board of directors or advisors and a Chief Executive Officer (CEO) plus Chief Financial

[3] Ceptureanu, Sebastian I., Eduard G. Ceptureanu, Cristian E. Luchian, and Iuliana Luchian. 2018. "Community Based Programs Sustainability. A Multidimensional Analysis of Sustainability Factors" *Sustainability* 10, no. 3: 870. https://doi.org/10.3390/su10030870

Officer (CFO) model with staff hired and managed within a traditional for-profit business structure. The Framework projects drew on community members for essential aspects of their decision making (e.g., via a membership-elected president) combined with some aspects of governance controlled and/or managed by paid personnel. This combination provided stability while increasing the likelihood of continued relevance. The paid staff (hopefully) have low turn-over rates and volunteer members provide a direct conduit between the project and their community. Database projects combined a paid staff with an advisory board drawn from the community. The governance of the Database projects is a kind of middle ground between the Middleware and Framework projects.

While the three types of projects used these three fairly distinct models of governance, it was clear that making their governance structure explicit and intentional was critical. Without formally defining the governance structure, none of the studied projects would have succeeded. Further, the governance structure in all the successful organizations needed to be able to change. It was also clear that choosing the correct governance structure for the project type was key. For example, the Middleware projects cannot be run as if they were Frameworks and vice-versa. This does not mean that Framework projects can never develop software – they can and often do – but Middleware projects are often in need of more top-down oversight to be able to achieve the development of more resource-intensive software.

4.1.4 Relevancy

For each of the projects in this study, the organization must strive to maintain relevance among its members, in order to be worth sustaining. How relevancy is maintained varies by organizational type. Middleware organizations need to understand and, to a degree, anticipate technological evolution. Database organizations can depend on the relative stability of scientific disciplines and a panel of experts to ensure their data holdings have the quality, attributes, and accessibility needed by their users. Framework organizations maintain relevancy primarily by incorporating members as major players in the governance process.

In all cases, the projects also ensured that they correctly tracked the project's initial objective, even though the means to obtain that objective might undergo significant change over time.

4.1.5 Uncertainty

Most projects faced periods of major uncertainty, often associated with funding and leadership issues. This period of uncertainty often occurred during the transition out of project initiation, which is often a very risky time for projects[4]. This is one aspect of the projects that is in complete alignment with challenges found in for-profit businesses.

[4] Skinner Katherine and Christina Drummond. 2022. The Quest for Home: Transforming from Grant-funded Project to Sustainable Operation. https://scholarlykitchen.sspnet.org/2022/05/19/the-quest-for-home-transforming-from-grant-funded-project-to-sustainable-operation/?informz=1&nbd=de50bbdb-968d-471a-ac98-0cfecba6bca0&nbd_source=informz

4.2 Assets

4.2.1 Collaboration and partnership

These qualities might make these projects unique in the world of business. While many businesses engage in both, all of the studied projects *prioritize* and put considerable resources into collaboration and partnership endeavors. In fact, most projects are not structured to stand completely on their own, and are instead part of tightly coupled networks, relying on partners for critical parts of their infrastructure.

For Middleware projects, partnership can take the shape of cooperative development where people working for one project do critical work for another project. In addition, the studied Middleware projects all leveraged externally developed open-source and commercial software to the extent that they could not exist without those connections.

For the Framework and Database projects, partnerships happened less on a technical level and more along community lines. The projects established by communities formed around a sense of belonging. Some of the Framework projects have formal members, while others do not. The Database projects generally have less formalized notions of membership. Regardless, both the Framework and Database projects depended on members (defined in the loose fashion) for essential aspects of their work. For example, Database projects must build up a large and relevant data store and all of those studied here did so by drawing on their community members to provide in-kind support. Without volunteers to provide data, Database projects would not work.

4.2.2 Value proposition

The most successful projects developed a growing group of motivated, engaged, and devoted participants, and had a clear value proposition for their community. Since Earth Science focuses on complex systems, community-based approaches (which includes collective thinking) are often appropriate for Earth Science projects.

Another aspect of the value proposition of Framework and Database projects is the opportunity for members of the community to have some level of ownership and control over the projects. This is true on several levels, including the increased ability to control the direction of critical infrastructure relevant to a member's field. In addition, there is a social component to membership and the deeper personal connections that intensified involvement in the project provides. In both cases, the value proposition of the project includes benefits that may be hard to define *a priori* for any given individual. Importantly, these study projects often provided a unique place for people and groups to gather, convene, and share ideas that are crucial for the progression of science and scientific technology.

4.2.3 Interoperability

Geological/Earth Science data can be so highly heterogeneous and specialized that “silos” often unintentionally form between domains. All three types of projects studied are, to some degree, working to provide interoperability as a way to break down those silos and remove barriers to interdisciplinary research. However, the types of projects diverge in their behavior. Middleware projects tended to focus on powerful computer- and information-science abstractions for common models that can straddle many disciplines. Framework projects tended to focus on social structures that bring together people from many backgrounds. Database projects tended to focus on building data stores of high quality information. This 'interoperability driver' is a major reason the three types of projects adopted different governance models. It is notable that for these projects, interoperability is an asset–not a problem to be solved. Also, interoperability can be solved by using both hard (computational) techniques and soft (human) approaches.

4.3 Lessons learned & Recommendations

In the above, we have identified key factors that influence sustainability, as identified through interviews. In this subsection we focus on the lessons learned from projects and provide recommendations.

4.3.1 Lessons learned

- There are two parts to sustainability: sustainability of the project; and sustainability of the product/mission. Recognition of this distinction will likely prove to be critical in moving forward with long-term planning to better support sustainability for projects.

- We noted the existence of three different types of digital data products: Digital databases (“Database”), Frameworks/Standards developments (“Framework”), and Middleware/software (Middleware). Identifying a typology (Middleware, Framework, and Database types) was helpful for determining some practices and pitfalls for sustainability.

- Most data infrastructure projects are developed by individuals or small groups of individuals with a shared vision in an “initiation phase”. Sustaining these projects requires development of a more formal governance structure in an “operational phase”. The attributes (skills, business models, staff, etc.) needed to be successful in the initiation phase are different from those needed at the operations phase of a project. The transition from the initiation phase to operations is often a major stumbling block. Almost all disciplinary scientists who were also project leaders were forced to “learn on the job” with respect to business models and community engagement.

- None of the studied projects began with a formal governance model. Instead, each project adopted a governance model over time. This approach worked because the identity, intentions, and community base was still unclear and evolving at initiation of a project.

- There is an inherent fragility associated with Database projects, which operate on unstable funding without the explicit backing of either major scientific societies or federal funding agencies. Most of the Database projects were addressing this issue, but data loss through project failure is a risk.

- Framework and Database projects spent significant resources on building community trust. Database projects in particular require engagement by trusted disciplinary scientists; their governance of database systems often included an advisory board made up of community volunteers. Framework projects were effectively inseparable from their community and delegated significant aspects of governance to that community.

- Middleware projects' governance often resembled a not-for-profit corporation. Middleware projects also worked on building community trust, but it was to a lesser degree and with less direct community involvement in the governance of the project, relative to Framework and Database project types.

- There is likely no “one size fits all” approach to sustainability within the Geosciences. Rather, success for the studied projects often relied on conditions that are hard to generalize and were specific to a specific project. Instead, the context and goals of each project determine the different types of approaches and business models needed to increase the potential for sustainability.

4.3.2 Recommendations

- US federal funding agencies should provide decade-long development support for projects to successfully progress through the initiation stage, particularly for science domain communities that do not yet have digital infrastructure.

- Projects should build community engagement into their initiation phase that is appropriate to the type of project. Projects should also be cognizant of the need to develop formal governance structures and a business model toward the end of an initiation phase.

- US federal funding agencies should create funding programs that more directly meet the needs of science data infrastructure. Data infrastructure projects are mission-driven, often addressing broader complex system issues. Projects must work on short-term time scales because of the nature of current research funding models, but all have the long-term burden of sustaining and supporting open science, interoperability, and community building. These additional long-term activities are often unfunded, because they do not create the products required by short-term funding.

- NSF should consider implementing a program similar to I-CORPS, but targeted at projects that aim to become (actually or effectively) not-for-profit organizations. Currently, there is little assistance to help projects develop as not-for-profit organizations. This could help bridge

the critical gap between project inception and the period when the initial leaders must first develop a formal governance mechanism.

- An NSF RCN-like project should be funded to bring together Database and Framework projects to discuss and innovate solutions for appropriate, long-term business models. These two project types require unique solutions because they inherently have a commitment to and contribution from a disciplinary community. Further, they are essential to current scientific practice. Thus, finding an appropriate model for a sustainable future for these types of projects would be highly beneficial.
- The Geoscience community needs to continue community-scale interactions (conversations, workshops, meetings, etc.) - in the absence of EarthCube - to pursue a sustainable future for data infrastructure.

5. Conclusions

This report presents the research design, approach, methods, data collection tools, results, and discussion points from a research study that aimed to examine the mechanisms used by Earth Science data infrastructure projects to achieve sustainability. Sustainability is a difficult goal to achieve and the scientific community does not yet know how to achieve it with respect to the newly emerging topic of digital data infrastructure. We have summarized the results of eleven interviews conducted with leaders of successful data infrastructure projects, with many in the Earth Sciences.

The projects - represented by the interviewees - were divided into three types: Middleware, Framework, and Database. We found significant structural differences among these three types of projects but they also faced some similar obstacles. All of the projects grappled with the need to establish governance and particularly the transition from their initiation phase to a more formalized operations phase. Similarly, all of the projects faced existential issues with funding and developed various ways to side-step the three-year research grant cycle, even though most of the projects were started using that funding mechanism. Lastly, each project's value proposition is closely tied to a community of users. Middleware and Framework projects tended toward a more diverse range of disciplines than the Database projects, where the focus on a specific field was more pronounced.

Taken as a whole, this study shows that while these three types of data infrastructure projects do vary in significant ways, their common struggles point toward ways NSF and other organizations could support their growth and sustainability.

Appendices

Appendix A: Methods

Approach

This study was designed as social science team research. Interviews with 11 key informants were used as the source of primary data. The study consisted of two phases with the first lasting ten months (February-November 2021) and the second lasting five months (January to May 2022); these phases were determined through oversight by the EarthCube office (the research funder).

Initiation of this project was the result of discussion during a meeting of the EarthCube Council of Funded Projects (CFP) led by co-authors Basil Tikoff (BT) and James Gallagher (JG)). The first phase of the project included its inception, formation of a 10-person research team selected by co-authors JG and BT, designing the research process and data collection templates, identifying the sample group, and interviewing leaders of five projects. Team members provided diverse perspectives and experiences relevant to understanding the broader context around the project; they had a spectrum of previous involvement in Earth Science including publishing, research, National labs, funded projects, and social science. A smaller core team of 3 individuals (co-authors Arika Virapongse (AV), JG, and BT) was established to increase the decision-making efficiency of the study, and provide overall coordination, research design, project management, and accountability for the project. Interviews were conducted by the core team.

For the second phase of the project, the research team was reduced to eight individuals. Outputs of this phase included six completed interviews, and analysis and write-up of the overall project.

Throughout the project, the research team met weekly or biweekly to develop and implement the project–often alternating between core group meetings (JG, BT, AV) and full team meetings (8 people total)–as we worked together to refine our process, gather information, and analyze data.

Since this project constituted human subject research, the interview process and questions were reviewed by the Institutional Review Board (IRB) at the University of Wisconsin Madison to ensure protection of the rights and welfare of humans participating as subjects in the research. In keeping with the intent of the IRB[5], we anonymized the interviews, to the extent possible, in our summaries of the interviews. Primary data (i.e., audio recordings, interview notes, and raw transcripts) are not available outside of the research group.

Data Collection and Analysis (Box 1)

To collect information about the nature of the projects, we conducted an approximately one-hour interview with a person who served, or currently serves, in an executive leadership role in a project.

[5]https://www.fda.gov/regulatory-information/search-fda-guidance-documents/institutional-review-boards-frequently-asked-questions

Each interview followed a script, question template, and general order (i.e., a semi-structured interview approach)(see Appendices). This approach provided consistency as well as flexibility so that (sub)topics relevant to a particular interview could be examined in more detail through follow-up questions. The interviews were recorded and automated transcripts were generated and saved.

Each interview was conducted by two or three core team members (AV, BT, and JG), with one taking the lead and asking questions while the other one or two listened and took notes; core team members rotated through the roles and all were present at most interviews to ensure consistency in how the interviews were conducted. The recording, transcript, and notes were made available to the full analytical team (10 in the first phase, 8 individuals in the second phase), who annotated the transcript (known as "coding" in qualitative analysis). In a subsequent meeting, the full team discussed their reflections about the interview, while one person from the team facilitated the discussion and took notes. The team then prepared a summary of the interview (consisting of quotes from the transcript, annotations, and notes from the facilitated discussion), and this served as a first-level derived product from the raw data of the interview itself. Most members of the team rotated through roles, so that the roles were distributed fairly evenly across the team.

During the coding and discussion process, we made sure that the questions of the project's sustainability model and why that model works were examined. Often information about these questions was intertwined with other aspects of the project and its operation. The number of team members who participated in the group coding and discussion of the interviews indicate the breadth of perspectives that were represented within the analysis.

Box 1. Summary of the Methods

Data collected/generated:

- Recording of the interview, resulting transcript[6]
- Notes by interviewers during the interview
- Publicly available background information about the project

Analysis process of the interviews:

- Coding by 3 to 6 (average 5) team members of the transcript to identify main themes
- Group discussion by 4 to 8 (average 5.9)[7] team members of the coded transcript
- Collection of all of the quotes, codes, and discussion notes into a summary document (first-level derived product)
- All of the coded quotes from all of the interviews were collected into a single document, and quotes were clustered together based on commonalities (second-level derived product)
- Clustered quotes were given a title and description, and these composed the final results

[6] Not publicly available because this information is protected by IRB requirements.
[7] One interview was not discussed by the group due to time/schedule constraints.

The summary document (first-level derived product) brought together quotes, insights about the quotes (coding), and notes from the facilitated discussion, so that we had all of the information about a given interview in one place. Next, all of the quotes and insights about the quotes from all of the summaries were gathered together. Then, the core team grouped this data according to common themes; these themes were given titles and descriptions (known as a clustering approach; second-level derived product). This final derived product formed the basis for the Results presented in this report.

We presented a first draft of this report to the eleven interviewees, and incorporated their feedback into the Discussion section. As a final step, we will use this report as a basis for a peer-reviewed paper, which will entail obtaining additional feedback from the interviewees and more in-depth analysis and discussion with team members.

Sample Group

The sample group was selected based on a two-tier process that focused on selecting a "project" first, and then an individual to interview as a representative of that project. For each of these tiers, the team determined criteria (Box 2), and then brainstormed projects/individuals that met the criteria.

Stratified sampling was used to increase the breadth of perspective included in the study (see criteria in Box 2), and these considerations were based on general knowledge within the research team as well as those that emerged as the team conducted interviews and identified patterns. In a scaffolding manner, the team first identified 2-3 initial sample group targets, completed the interviews, and completed the data collection summaries. Once that was completed, the team then focused on identifying the next few sample group targets. In this way, the team sought to reduce redundancy and fill in gaps in knowledge for the overall study.

Box 2. Sample group criteria

Project sample group criteria:

- Relevant to Earth Science data
- Official location in the US
- Existed for 10+ years
- Not a government-based project or national labs
- Stratified sampling considerations:
    - Sub-sample group designation
    - Size of the project (number of staff)

Individual sample group criteria:

- Had/has a strategic role in the program
- Held their leadership role for at least 2 years
- Stratified sampling considerations:
    - Gender
    - Career stage

Table 1 shows the projects that were included in the study, and some key characteristics that were relevant to our sampling approach. The study projects were selected to represent sub-sample groups of Database (5 projects), Middleware (3), and Framework (3). Database projects aimed to bring together data and data resources for use. Middleware projects sought to develop software and technology. Framework projects focused on developing best practices and guidelines. The identity of the individuals interviewed in the study are not revealed per our informed consent agreement and IRB requirements for exemption.

Table 1. List of Projects. Projects included in the study and their defining characteristics.

Project name & Website	Sub-sample group	Organizational structure	Year founded	Current staff
BCO-DMO www.bco-dmo.org	Database	Univ. hosted, NSF funded	2006	5
ESIP esipfed.org/	Framework	501(c)3	1998	5
Force11 force11.org	Framework	501(c)3	2011	16
HDF Group www.hdfgroup.org	Middleware	501(c)3	2006, NCSA 1988	20

IEDA www.iedadata.org	Database	Univ hosted, NSF funded	2010 (based on the web site copyright)	14
IRIS www.iris.edu/hq/	Database	NSF funded	1984	50
OGC www.ogc.org/	Framework	501(c)3	1994	20
OPeNDAP www.opendap.org	Middleware	501(c)3	2000, Univ of RI 1993	5
PBDB paleobiodb.org	Database	NSF Funded	1998	3
SERC serc.carleton.edu/	Database	Univ. hosted, NSF funded	2001	19
Unidata www.unidata.ucar.edu/	MIddleware	UCAR Hosted, NSF funded	1984	~20

The demographic data of the individuals in the sample group were all self-reported. The sample group consisted of 6 women and 5 men. When asked if they felt that they were part of an underrepresented group within the Earth Sciences and Data Sciences, interviewees identified the following characteristics about themselves: not being a US citizen (n=2), “Out and proud gay man”, ”not having a PhD”, “I’m a neuroscientist”, “I grew up in a rural environment”, and being a professional musician.

Interviewees reported having these terminal roles in their project: Director (n=3), President (n=2), CEO (n=2), Funding program manager, Lead, Senior research scientist, Executive Director; all were paid positions with the exception of Lead. The years that interviewees have been involved in their project ranged from 5 to 30 years (average 18 years). When they started working on the project, interviewees were in these career stages: Early/Middle (n=3), Middle (n=3), and Middle/Late (n=3), and Early (n=2).

Appendix B: Communication with interviewees

Initial contact with interviewee

Hello [individual’s name],

I’m reaching out on behalf of a research team working on an NSF-EarthCube office funded project (on behalf of the Council for Funded Projects) that is looking at Sustainability of Earth Science Data Infrastructure. As the EarthCube program enters into a final year of funding, it is exploring ideas around how to improve the sustainability of Earth Science data projects. Our research entails

interviewing individuals from selected projects to hear their perspectives and experiences around the sustainability of the project. The outcome of the project will consist of recommendations made to EarthCube.

[Project] was mentioned as a project that we might want to explore. A member on our research team suggested that I reach out to you to ask if you might have a recommendation of someone to interview. This person could also be you! We are looking for someone who has held or currently holds a leadership role in [Project], and has been involved in making major decisions about the project.

We would be grateful for any suggestions or assistance in connecting with someone. If you are interested in being interviewed, I'd be happy to provide more details about the project.

Thank you!

Scheduling the interview

Dear [individual],

Thank you so much for agreeing to participate in this interview.

Here are more details about the interview:
The interview will be scheduled for 1 hour, conducted via Zoom with video, and recorded. A transcript of the interview will be used for analysis purposes. We do offer a $400 honorarium for interviewees with the intention of off-setting any costs (opportunity costs or otherwise) associated with participating.

Please take a look at the informed consent (link provided), which describes the study in more detail and what you should expect as a participant. The study protocol was reviewed for human subjects research ethics by the Institutional Review Board at University of Wisconsin and determined exempt. Before we begin the interview, you'll have an opportunity to ask questions about the project, and we'll formally ask for your consent to participate.

We realize that our community is pretty tightly knit, so we'd like to let you know who is involved in this project, and who would be aware of your identity. The project leads are: James Gallagher, Basil Tikoff, and Arika Virapongse (the three of us conduct the interviews). The rest of the project team is made up of: Chad Trabant, Peter Cornillon, Rebecca Koskela, Brooks Hansen, and Susan Shingledecker. In addition, administrators from the EarthCube office who would be handling the honorarium would also be aware of your participation in the project. **Please let us know if you feel that you have a conflict of interest and would prefer NOT to continue to participate in an interview.**

If you would still like to continue in the project, we'd like to schedule a time for us to do the interview. Would you be available for 1 hour during the month of [month] at these times of the week [specific times offered].

If you can select a date/time, I can send around a google calendar event to confirm.

Appendix C: Interview script and template

***Script for start of interview:* (*5 min*)**

"Thank you for agreeing to be interviewed for our study entitled "Sustainability models for integrated digital Earth Science beyond EarthCube" that seeks to answer the question: "**What organizational/business models do Earth Science groups use to support their sustainability?"** As described in the informed consent, this project is being conducted by the Council of Funded Projects (CFP) of EarthCube and is funded by the EarthCube program of the US National Science Foundation.

[Introduce the interviewers, including their role in the project and their role in the interview].

We hope that you've had a chance to read the informed consent in detail. At any time during this interview, you should feel free to decline to respond to a specific question (you can say something like "no comment"), and stop the interview at any time. You can also ask us to retract all of your responses from being included in the study after the interview is done, as long as the study has not been published.

In terms of your privacy and anonymity in the study, we will be reporting a synthesis of the results from the interviews that we conduct. In the report, we will reveal the name of the organization that you discuss, but not your name. However, due to the specific nature of your role in the organization, it might be possible for someone to deduce your identity. Your name will be revealed only within the analysis team (a team of about 10 individuals who are working on this project).

In this interview, we will be asking you some demographic questions about yourself and the organization that you worked/work for. Then, we'll be asking you some questions about your perspectives around the organization and your work there. Finally, we might reach out to you with some follow up questions, including providing a draft of the final product for you to comment on. We would also like to note that we designed our interview questions to only collect information that we plan and hope to use in the analysis.

We expect that this interview will take approximately 1 hour to complete. We would also like to record this interview for analytical purposes.

Do you have any questions about the study?

Do you agree to take part in this study?
Do you confirm that you have read and understand the informed consent?
[Take a note of these responses]

In that case, we will now begin the interview."

Interview questions
[secondary bullet points are follow-up questions]

Demographics of the interviewee (*10 min*)

- Name of the interviewee
- Diversity, equity, and inclusion are important to us. What is your gender identity?
- Are there any other characteristics about yourself that you'd like to share, especially if you feel that you are part of an underrepresented group within Earth Sciences and data sciences?
- Could you share a brief 1-2 minute summary about your professional background that is relevant to your role(s) at the project that you'll talk about, including your education & other work experience?

Relationship of the interviewee to the project (*10 min*)

- What is the name of the project that you will talk about today?
- What Earth Science domains does the project work with?
- When was the project started?
- What year did you get involved in the project?
- What roles did/do you serve in the project?
    - How many years did you serve in each role?
    - What career stage(s) were you in when you were in that role(s)? For example, early, middle, late.

Description of the project: (20 min)

- Can you describe the organizational structure of the project?
    - How geographically distributed is the project?
    - How are decisions made in the project?
    - Who makes decisions?
    - What does the decision-making process look like?
    - How are the end users (beneficiaries) of the project involved in making decisions?
- How would you describe the people involved in the project?
    - What is the educational level of the people involved?
    - What are different roles that people take in the project (e.g., staff, in-kind support, volunteer leaders etc)?
    - How diverse is the group of people that work on the project?
    - What approaches were used to increase the diversity of the group working on the project?

    - How many staff are directly employed by the project?
    - How many volunteers do you have serving the project?
    - How distributed are the members of your project?
- Can you describe the end-user community that is associated with your project?
    - How would you define your community?
    - How large is the community that the project serves?
    - How many people are in your community?
    - What makes someone a member of your community?
    - What motivates a person to be involved in your community?
    - How do you engage with your community?
    - How is your community involved in decision making in your project?
- What are the intended benefits of the project?
    - What innovations, services, or products did the group intend to improve?
    - How did/does the project benefit the people they serve?
    - What do people get from your project?
    - How does the project maintain its relevance to its beneficiaries? (users, members, participants)

Business model & Sustainability: (15 min)

- Can you describe the business model of the project?
    - How is your project funded?
    - Has the business model changed at any time? What motivated that shift? What were the challenges associated with that shift?
    - Are you in the process of trying something new or thinking about trying something new in your project? If so, what is driving that decision?
    - What funding did your project explore that didn't work out?
- What does sustainability mean for the project?
    - Using your definition of sustainability, how sustainable do you think the project is?
    - How important is sustainability to the project?
    - What are challenges to the sustainability of the project?
    - What do you think would increase the sustainability of the project?
    - What are some external factors that impacted your sustainability? External factors can be things, events, or changes, for example, funding, competition, shift in funding agency priorities. Were those factors foreseen? How was your sustainability impacted as a result of those factors?
    - Were there any events that led to more stability to the group?
    - Were there any events that led to instability to the group?
    - Can you describe any moments when the group got close to failing?
- What would long-term sustainable success look like for your project (blue skies)?

Closing (5 min)

Those are all of the questions that we have for you today. Is there anything else that you'd like to add? Or do you have any questions that you'd like to ask?

In terms of next steps, we'll be working on the analysis of these interviews over the next month. Then, we have a follow up activity that is voluntary, but we would also ask for your participation. We would like to invite you to read and provide commentary on the internal EarthCube report. The activity is designed to allow you to correct factual errors or misunderstandings in how your organization operates. So, we'll reach out to you again soon.

Thank you so much for your time. If you have any questions about this project later, please do let us know.

Appendix D: Informed Consent

We are asking you to participate in a research study titled "Interviewing professionals who developed successful cyberinfrastructure in the Earth Sciences". This study is being led by Basil Tikoff (University of Wisconsin – Madison), James Gallagher (OPeNDAP), and Arika Virapongse (Middle Path EcoSolutions/Ronin Institute). The study is funded by the National EarthCube Office, which is funded by the National Science Foundation.

What the study is about
The purpose of this research is to address core issues of project sustainability relevant to EarthCube by examining similar projects that have demonstrated success by outlasting their initial funding and period of inception. We seek to invite a select group of individuals, who have led successful Earth Science projects from or near their beginning to a point where they are clearly self-sustaining, to discuss the strategies and tactics they employed as leaders of the project. We will choose projects and people who are closely tied to the EarthCube community and thus who are likely to understand the issues it is facing.

What we will ask you to do
We will ask you to answer questions about yourself, how you got involved in the digital Earth Science infrastructure, and details of the organization that you are involved in. We anticipate that the interview will take 60 minutes, but may last up to 80 minutes. The interview will take place by Zoom.

Risks and discomforts
We do not anticipate any risks from participating in this research.

Benefits
We anticipate that the benefits from this research will be to produce and provide helpful information about how to make Earth Science cyberinfrastructure projects more sustainable.

Compensation for participation
You will be compensated $400 to provide us with an interview, as reimbursement for your time spent participating in the interview and for any expenses that you might incur during the process.

There is one follow up activity that is voluntary. We will ask you to proofread and provide comments on the internal EarthCube report prior to submission. This follow-up activity is designed to allow you to correct factual errors or misunderstandings in how your organization operates.

Video Recording
We will be making a video recording of your interview. Agreeing to the consent form is an acknowledgement that this will be done.

Privacy/Confidentiality/Data Security
Because of the highly individualized nature of your work, we cannot guarantee confidentiality. In the written report, we will not identify you by name.

We anticipate that your participation in this survey will present no great risk to you as a professional and individual.

Please note that email communication is neither private nor secure. Though we are taking precautions to protect your privacy, you should be aware that information sent through email could be read by a third party.

Your confidentiality will be kept to the degree permitted by the technology being used. We cannot guarantee against interception of data sent via the internet by third parties.

Taking part is voluntary
Your involvement is voluntary. You may refuse to participate before the study begins, discontinue at any time, or skip any questions/procedures that may make you feel uncomfortable. If you choose not to participate prior to the interview, you will not receive monetary compensation.

Follow up studies
We may contact you again to request your participation in a follow up study. As always, your participation in any follow-up activities will be voluntary.

If you have questions
This interview and study is done as a group effort. However, the lead researcher conducting this study is Basil Tikoff, a professor at the University of Wisconsin-Madison. If you have any questions, you may contact Basil Tikoff at basil@geology.wisc.edu or at 608-262-4678. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) for Human Participants at (608) 262-9710 or access their website at https://research.wisc.edu/compliance-policy/human-research-protection-program/education-and-social-behavioral-science-irb/.

Statement of Consent
You will be asked at the beginning of your interview if you consent to this activity.