Middle Path
EcoSolutions

OPeNDAP

How Earth Science Data Infrastructure Projects Become Sustainable

Arika Virapongse, James Gallagher, Basil Tikoff

With gratitude to team members: Peter Cornillon, Rebecca Koskela, Susan Shingledecker, Chad Trabant; Brooks Hanson

Presentation for:

"Sustainable, Equitable and Usable Earth Science Cyberinfrastructure", session at GSA Connects (Annual) meeting, Anaheim, CA, September 24, 2024

How this all got started

- **EarthCube founded the Council of Funded Projects**
- ***The EarthCube Council of Funded Projects was purposefully non-equal member of the EarthCube hierarchy (note: this was the participants' choice, not that of leadership)***
- **Funding (minor) became available for different groups in EarthCube to achieve specific goals. The members of the Council of Funded Projects wanted to know: How can we sustain the digital products that we are producing?**
- **As chair of EarthCube Council of Funded Projects, I requested funds. Arika and James joined the effort.**

EARTHCUBE GOVERNANCE

The EarthCube governing structure[2] evolved over the life of the program. The final structure consisted of six elected bodies:

- Leadership Council (LC) - the elected voice of the EarthCube community, establishing strategic direction for the program and making decisions critical to success.
- Science and Engagement Team (SET) - a connection between the academic geoscience and technology communities in EarthCube that linked EarthCube activities to relevant organizations and initiatives.
- Technology and Architecture Committee (TAC) - a forum for maintaining an architecturally-oriented overview of EarthCube's technological capabilities.
- Council of Funded Projects (CFP) - a forum for project personnel to interact, discover and work together on common needs.
- Council of Data Facilities (CDF) - a federation of existing and emerging geoscience data facilities exchanging experiences and promoting standards and best practices in the organization and operation of data facilities.
- Nominations Committee - a body that oversees the nomination of EarthCube community members to various Governance roles.

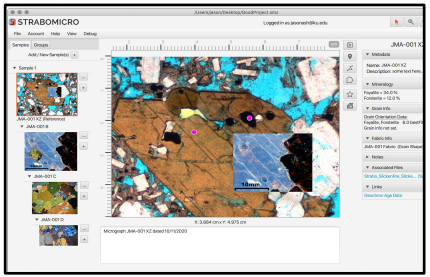
A personal perspective

Why was I motivated?

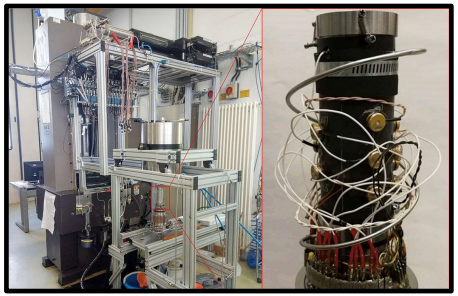
A requiem for EarthCube

StraboField

StraboMicro



StraboExperimental



What we did

Time frame: Feb 2021 - Feb 2023

Process:

1. Formed a 10-person research team
2. Developed questions to ask interviewees
3. Obtained IRB exemption approval (seems like a small, innocent task...)
4. Conducted semi-structured interviews (~1 hour long) with representatives from 11 projects; done by the core group of 3 people
5. Conducted more detailed qualitative analysis with research team
6. Wrote a report for EarthCube with research team
7. Rewrote the report as a scientific manuscript and submitted to a peer-reviewed journal with core group

What we did

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 leadership role in the program
- Held their leadership role for at least 2 years
- Stratified sampling considerations: gender & career stage

Study Projects
BCO-DMO www.bco-dmo.org
ESIP esipfed.org/
Force11 force11.org
HDF Group www.hdfgroup.org
IEDA www.iedadata.org
IRIS www.iris.edu/hq/
OGC www.ogc.org/
OPeNDAP www.opendap.org
PaleoDB paleobiodb.org
SERC serc.carleton.edu/
Unidata www.unidata.ucar.edu/

Quick definitions

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.

PROJECT NAME & WEBSITE	SUB-SAMPLE GROUP	ORGANIZATIONAL STRUCTURE	YEAR FOUNDED	CURRENT STAFF
BCO-DMO Biological & Chemical Oceanography Data Management Office www.bco-dmo.org	Database	University hosted, NSF funded	2006	5
ESIP Earth Science Information Partners esipfed.org/	Framework	501(c)3	1998	5
Force11 Future of Research Communications and e-Scholarship force11.org	Framework	501(c)3	2011	16
HDF Group Hierarchical Data Format Group www.hdfgroup.org	Middleware	501(c)3	2006, NCSA 1988	20
IEDA Interdisciplinary Earth Data Alliance www.iedadata.org	Database	University hosted, NSF funded	2010 (web site copyright)	14

PROJECT NAME & WEBSITE	SUB-SAMPLE GROUP	ORGANIZATIONAL STRUCTURE	YEAR FOUNDED	CURRENT STAFF
IRIS Incorporated Research Institutions for Seismology www.iris.edu/hq/	Database	NSF funded	1984	50
OGC Open Geospatial Consortium www.ogc.org/	Framework	501(c)3	1994	20
OPeNDAP Open-source Project for a Network Data Access Protocol www.opendap.org	Middleware	501(c)3	2000, University of Rhode Island 1993	5
PaleoDB Paleobiology Database paleobiodb.org	Database	NSF Funded	1998	3
SERC Science Education Resource Center serc.carleton.edu/	Database	University hosted, NSF funded	2001	19
Unidata www.unidata.ucar.edu/	MIddleware	UCAR hosted, NSF funded	1984	20

Some questions we asked

Demographics of the interviewee

Relationship of the interviewee to the project

Description of the project

- How long has the project been around?
- What Earth Science domains does the project work with?
- Can you describe the organizational structure of the project? (who does what? How do things get done?)
- Can you describe the end-user community that is associated with your project?
- What are the intended benefits of the project?
- What does the decision-making process look like?

Business model & Sustainability

- How is it funded?
- How has the business model changed over time?
- What does sustainability mean for the project?
- What are challenges to the sustainability of the project?
- Can you describe any moments when the project got close to failing?
- What would long-term sustainable success look like for your project (blue skies)?

Data collected/generated

- Recording of the interview, resulting transcript (not publicly available because this information is protected by IRB requirements)

- Notes by interviewers during the interview

- Publicly available background information about the project

Analysis Process of the Interviews

- Coding of the transcript by 3 to 6 (average 5) team members, to identify main themes

- Group discussion of the coded transcript by 4 to 8 (average 5.9; one interview was not discussed by the group due to time/schedule constraints) team members

- 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

Results (1 of 4)

There are Middleware, Framework, and Database project types. There were significant structural differences among Middleware, Framework, and Database projects but they also faced similar obstacles.

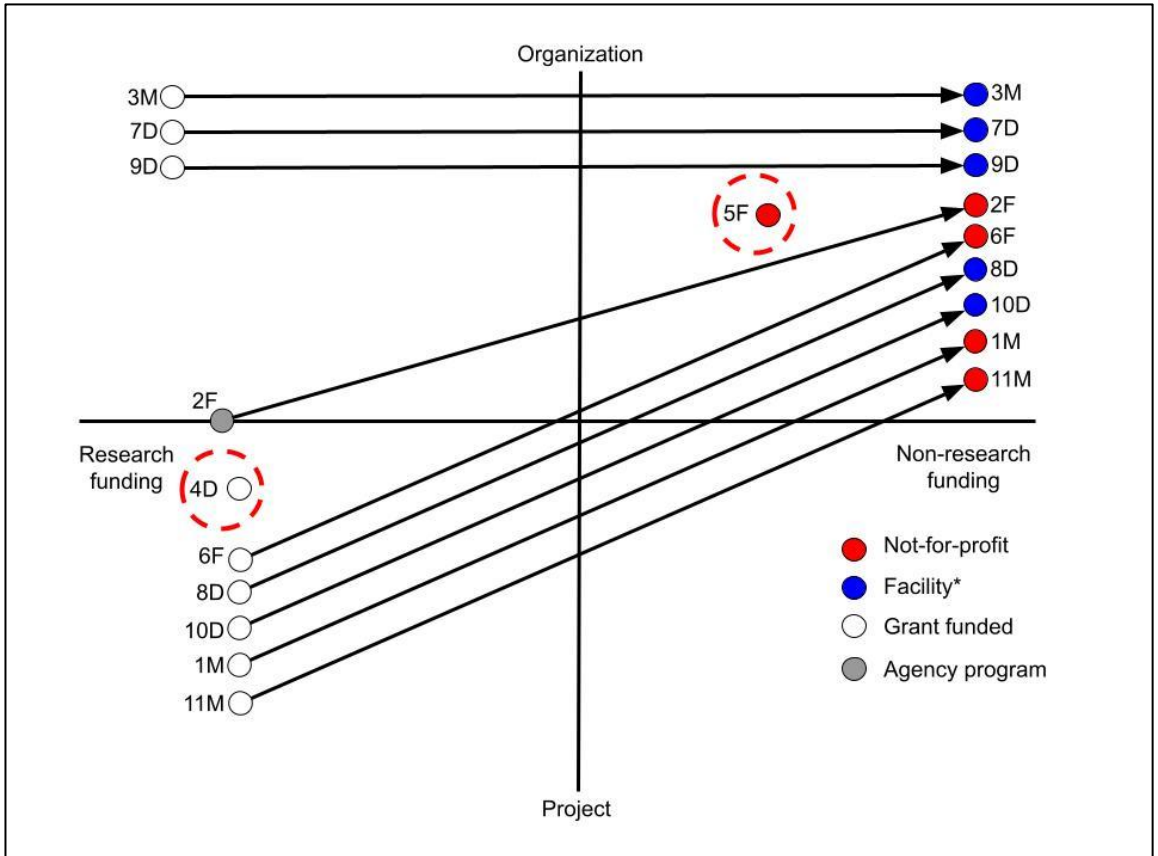
Leadership evolution. For projects that are science-driven, practicing scientists play major leadership roles in the initial stages of the projects. As projects matured, different types of leadership were sometimes needed; for example, leaders skilled with building communities were needed if a project depended on community engagement. The most successful projects were able to identify the right leadership at the right time in the project.

Results (2 of 4)

Flexible governance structure. 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 were still unclear and evolving at initiation of the projects.

Projects that do not transition to organizations are most at-risk. Part of the definition of a project is that it remains in a short-funding cycle with no long-term business model. 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.

Results: A graphic form



Results (3 of 4)

A project's value is closely tied to their 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. 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. The most successful projects developed a growing group of motivated, engaged, and devoted participants, and had a clear value proposition for their community. All projects began with an innovative idea and/or critical development that fulfilled user needs.

Results (4 of 4)

Middleware projects are outliers in science. Middleware projects' governance often resembled a for-profit corporation.

Research funding is a poor fit for projects once they become organizations. All of the projects faced existential issues with funding and developed various ways to sidestep the three-year research grant cycle, even though most of the projects were initiated using that funding mechanism. Despite addressing issues that are essential for science today, the projects often have long-term goals that the research grants were not originally designed to support. Most projects faced periods of major uncertainty, often associated with funding, because there is no clear path for continued funding for digital infrastructure in the venues provided to scientists through governmental agencies. Each project spent significant effort finding ways to fund their digital initiatives

References/products

Virapongse, A.; Gallagher, J.; Tikoff, B.; Cornillon, P.; Koskela, R.; Shingledecker, S.; Trabant, C.; Hanson, B. (2022). Sustainability models for integrated digital Earth Science. In EarthCube Organization Materials. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0JH3MBN

EarthCube 2023 Poster: https://drive.google.com/file/d/160UfA97eE9_F_8g6NrlhGmmq-dC4smFL/view?usp=sharing

Virapongse, A.; Gallagher, J.; and Tikoff, B. (2024) Insights on Sustainability of Earth Science Data Infrastructure Projects. Data Science Journal, v. 23, pp. 1–27. doi: https://doi.org/10.5334/dsj-2024-014.

Insights on Sustainability of Earth Science Data Infrastructure Projects

RESEARCH PAPER

ARIKA VIRAPONGSE
JAMES GALLAGHER
BASIL TIKOFF
*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

We studied 11 long-term data infrastructure projects, most of which focused on the Earth Sciences, to understand characteristics that contributed to their project sustainability. Among our sample group, we noted the existence of three different types of project groupings: Database, Framework, and Middleware. Most efforts started as federally funded research projects, and our results show that nearly all became organizations in order to become sustainable. Projects were often funded for short time scales but had the long-term burden of sustaining and supporting open science, interoperability, and community building-activities that are difficult to fund directly. This transition from 'project' to 'organization' was challenging for most efforts, especially in regard to leadership change and funding issues.

Some common approaches to sustainability were identified within each project grouping. Framework and Database projects both relied heavily on the commitment to, and contribution from, a disciplinary community. Framework projects often used bottom-up governance approaches to maintain the active participation and interest of their community. Database projects succeeded when they were able to position themselves as part of the core workflow for disciplinary-specific scientific research. Middleware projects borrowed heavily from sustainability models used by software companies, while maintaining strong scientific partnerships. Cyberinfrastructure for science requires considerable resources to develop and sustain itself, and much of these resources are provided through in-kind support from academics, researchers, and their institutes. It is imperative that more work is done to find appropriate models that help sustain key data infrastructure for Earth Science over the long-term.

Four concrete take-aways

Many disciplinary communities are underserved – or completely unserved – in terms of digital needs.

We do not yet have any model for sustainability for grass-roots digital products. Many digital products, particularly in Geology, have no pathway to become part of an organization.

Different types of leadership are generally needed as a database project transitioned to an organization. Organization leaders need to be skilled with building communities.

Funding agencies have not adjusted to the reality that digital products are essential to current research; neither have many disciplinary science communities.

Before you ask, let me try to answer…

Many people have this basic question: How can these groups that seem to exist outside any formal institution survive?

-Pure stubbornness (can be cast as perseverance)
-They have diverse strategies for earning income
-Some in-kind support (e.g., from universities)