Author: James Colliander (colliand@2i2c.org)
Affiliation: 2i2c.org & UBC
NASA: SMD RFI NNH23ZDA005L (Aspect 1)
Fiscal sponsor: Code for Science and Society

Building the Open Source Science Stack

Science is a social enterprise that builds and organizes human understanding of the universe with testable explanations and predictions. Science is a doorway through which we can walk into a better tomorrow[1]. The enterprise of science, carried out in social structures like universities and colleges, government agencies, academic societies, conferences and workshops, philanthropic institutions, publishers, laboratories, research teams,..., constantly evolves. The unification of federal agencies[2] to declare 2023 as the *Year of Open Science* is an historical opportunity to innovate and improve science at NASA and beyond.

The inputs of the scientific enterprise are people and resources organized in social structures. The people *participate* in science by working with the resources and each other to generate the outputs of science: advances to human understanding of nature. What would it take to boost scientific productivity by an order of magnitude in a decade? A linear answer: a 10X boost in the inputs – people and resources – will generate a 10X boost in the outputs. The more exciting answer is nonlinear: improvements to the ways people participate in the science enterprise – *open* science – will generate massive gains in productivity.

Three guiding ideas summarize 2i2c's advice for NASA. We encourage NASA to:

1. Recognize each **person's right to participate** as a core tenet of *open* science
2. **Eliminate accidental complexity** to enable participation and collaboration
3. **Build "digital villages" enabling open science communities** to flourish.

2i2c will collaborate with NASA to enable communities that want to do science in the open by empowering them to do so with state-of-the-art digital and social infrastructure. These ideas will be developed further using in-line responses to prompts in Aspect 1.

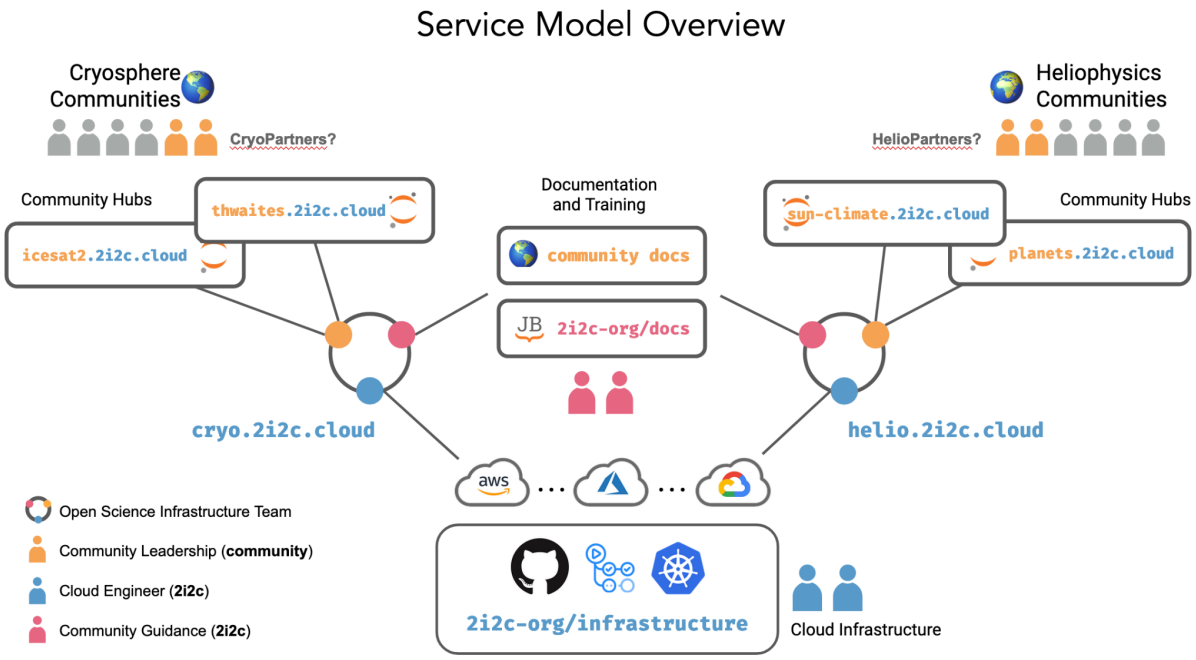
2i2c's experience with SMD data and computing resources

2i2c designs, deploys and operates interactive computing platforms for a variety of communities that use SMD scientific data. These communities include Pangeo, CryoCloud, Cooperative Institute for Research to Operations in Hydrology (CIROH), MyBinder, NASA VEDA, Multiscale Machine Learning In Coupled Earth System Modeling (M^2Lines), Learning the Earth with Artificial Intelligence and Physics (LEAP), Jupyter Meets the Earth (JMTE), Jack Eddy Symposium[3], [4], OceanHackWeek, GridSST Hack Week, Volcanology Hub for Interdisciplinary Collaboration, Tools and Resources (VICTOR), OpenScapes, LinkedEarth, EarthLab, CarbonPlan. These communities use data from SMD and other sources to investigate diverse phenomena in the Earth system. 2i2c also works[5] with The Carpentries, MetaDocencia, Center for Scientific Collaboration and Community Engagement, Open Life Science, Invest in Open Infrastructure, Chan-Zuckerberg Initiative, Callysto, CloudBank, Berkeley's Computing, Data Science and Society Division, GESIS, Curvenote and many colleges and universities to broaden and improve participation in open source science.

2i2c's team has collaborated with several of these communities over many years (including before 2i2c existed). Our team contributes by **managing cloud infrastructure** for interactive computing, collaborating with scientists on **open source development** to facilitate open science workflows, and providing **guidance and support** in using this infrastructure. The infrastructure 2i2c operates for these communities is used for a variety of use-cases including training, exploratory data analysis, deep research, machine learning, and collaborative authoring of research reports. These communities use a common open source scaffold deployed on commercial cloud to deliver their own curated end-user software environments. By delivering these platforms interoperably across the commercial cloud, 2i2c's approach avoids scenarios where a vendor controls science infrastructure and protects the right to participate in science.

As the infrastructure operator with visibility across these communities at once, 2i2c identifies interesting patterns. Each is a kind of **"digital village"** that facilitates access to data, computation and software around a shared focus area. These communities:

- are **cross-organizational**: members do not belong to a single university or industry
- are **multi-stakeholder**: students, researchers, people from industry, various branches of government, and others join these communities for diverse purposes
- **use the same open source infrastructure** with customization for their use-cases
- **eliminate accidental complexity** using a shared responsibility model
- **produce and share knowledge in many different forms** such as documentation, data, dashboards, repositories, interactive books[6], and peer-reviewed articles.

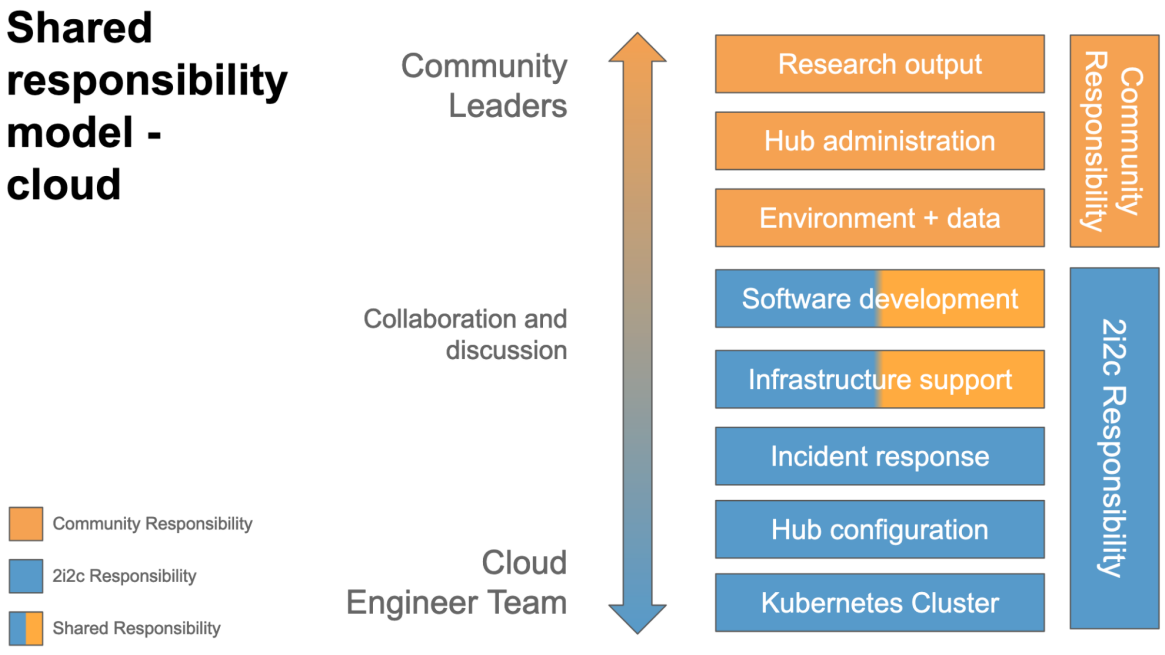


We believe that communities like this are the future of open science[7]. It is critical to understand the technical and social conditions that led to their success, and how we can try to replicate them elsewhere.

Pangeo, a showcase example for open source science, has inspired many communities. In summer 2016, a group of **scientists and software developers** began imagining a single "standard" toolkit for climate-related science, similar to what astropy had become for

astronomy-related sciences. The group gathered for a workshop in October 2016 that was socially engineered with guidelines for effective presentations and a show-more-than-tell call to action. That workshop was the launch event for Pangeo, a digital village that grew beyond the initial motivation to build tools. Pangeo's emergent success can partly be attributed to the potent mixture of skills among workshop participants, and the strong focus on solving tangible problems faced by the attendees. Amplifying their disciplinary expertise in climate science, the Pangeo launch team also had the software and commercial cloud skills to eliminate accidental complexity and overcome obstacles to setting up cloud-native workflows. Pangeo evolved to reach the limits of its technical capacity to manage complex cloud infrastructure. 2i2c was created in-part to robustly operate this infrastructure for Pangeo, and to provide the commercial cloud expertise necessary to replicate Pangeo's success for other research communities.

What can we learn from the initial conditions that led to Pangeo? A multi-disciplinary group of scientists and open source software developers worked together to address common challenges in their research area. Participants were encouraged to be **intellectually generous**, document their problems and solutions, and share their work so that others could build atop their progress. We recommend NASA recognize that many teams that wish to emulate Pangeo's success are blocked by the accidental complexity[8] in setting up cloud-native workflows. We encourage NASA to **fund efforts to learn from successful open source science communities** using ethnography or other social research methods.



SMD's infrastructure that successfully supports open science

2i2c commends NASA for catalyzing a global **transformation toward open source science**. TOPS, SPD41a, OSSI, and other efforts by NASA enabled OSTP[2] to unify federal agencies around a common vision for the *Year of Open Science*. 2i2c's network of international partners are inspired by NASA and are advocating for aligned open science efforts within their countries. We especially appreciate NASA's focus on culture, human

resources, and open source software as critical components for the open science transformation.

NASA awards (TWSC22; IceSAT2 Science Team) support CryoCloud, a social + technical experiment aimed at opening up cryosphere science. Through this project, a community of investigators has persistent access to the CryoCloud JupyterHub (operated by 2i2c). The hub and collaborative workflows will be iteratively improved through a sequence of training events, conferences and workshops culminating in a special session demonstrating the benefits of open source science at AGU 2023. We encourage NASA to **launch other social + technical experiments to identify effective ways to advance open source science.**

NASA's systems science and research-to-operations expertise are vital for open science communities. SMD's Living with a Star program unified stakeholders with apparently disparate interests (stellar dynamics, magnetohydrodynamics, solar energetic particles, space weather,...) in the Earth-Sun system.The Cooperative Institute for Research to Operations in Hydrology (CIROH), a national consortium focused on management of water resources, is similar. CIROH unifies stakeholders with diverse motivations – hydrology research, flood preparedness, managing water for irrigation and hydroelectric power – using digital infrastructure (operated by 2i2c) and forecasts from the National Water Model. WIFIRE, another multistakeholder "digital village", uses cloud-native data/compute, to manage wildfire. A similar "digital village" could manage risk and improve resiliency for space weather events[9]. Advances in data, software, and computing are enabling transformational, interdisciplinary science, changing the realm of possible questions[7]. We encourage NASA to identify themes and innovations to **create flourishing multi-stakeholder science + software communities to advance understanding and generate societal benefits** from SMD data.

SMD's infrastructure that hinders open science

Social innovation is under-appreciated when designing digital infrastructure. Pangeo's transition to cloud-native workflows, a technology move that eliminated accidental complexity, enabled new ways to collaborate on essential problems in climate science. Elimination of accidental complexities in *software* will not massively boost the productivity[8] of a single engineer or scientists. Removing complexities that block *collaboration* unleashes the collective human effort toward finding solutions to grand challenge problems! We encourage NASA to expand focus from "current scientific data and computing" to include the social infrastructure (governance, mission, values, codes of conduct) required for vibrant open source science communities. **Social innovation is the magic for open science.**

Prior art and training for other communities to follow is lacking. The "digital villages" served presently by 2i2c present a repeatable pattern that opens science. The technical infrastructure – an open source elastically scalable interactive computing platform with a curated toolchain deployed on commercial cloud and adjacent to a "data watering hole" – efficiently supports training, multi-stakeholder collaboration, exploratory analysis, and sharing of computational narratives. 2i2c's approach to delivering the technical infrastructure guarantees the community's right to replicate with interoperability across the commercial cloud vendors. The **right to replicate** and protections against vendor lock-in flow from the

individual's **right to participate** in open science. **Open source science infrastructure should be a public good**. NASA should **build a printing press for "digital villages",** a timely[7] innovation to support communities wishing to do science in the open. The village metaphor provides context for scoping an individual's right to participate in science.

Institutional gatekeeping makes science inaccessible to under-privileged communities. NASA funded technology platforms deployed by institutions often favor scientists at those institutions and are inaccessible, less accessible, or inconveniently accessible to those outside of the institution. NASA should develop a **right to participate in open science** and design technical guidelines for appropriate open access to infrastructure used by open science communities. NASA should **support services that provide wide access that is not restricted to those with institutional accounts**.

Most data is not in a cloud-native format, and is thus inaccessible in the cloud. Significant effort is required to do the necessary transformations and movement to make large datasets more accessible. The Pangeo Forge project was created to help address this problem. NASA should **fund development and standardization efforts to facilitate the migration of NASA data into a cloud-native and vendor-agnostic form.**

Reward systems in funding encourage gatekeeping and isolation. A transformation to open science will involve changes to all aspects of the science process. The ways scientists expose data, explore ideas, collaborate, author[6] documents, submit for publication, submit for grants, carry out peer review, publish results, read literature, build upon other's work, and train junior colleagues will all change. Complex grant competitions implicitly favor proponents from well-resourced institutions. Competition between universities for research rankings hinder the formation of multi-stakeholder open science teams. There is a tension between what's better for science versus what's better for my institution; what's better for my research community versus what's better for my lab; what's better for my trainees versus what's better for the next cohort of trainees in my discipline. Competition among principal investigators hinders science team formation. Digital and social infrastructure should be designed so that the easiest path to use these resources aligns with best practices for open science, broadening participation, etc. We recommend that NASA **design funding calls that reward proposals with a coherent collaboration model**.

Technology is replicated many times in silos rather than built once together. Too many groups within NASA identify their problem, build their own solution, and then share their source code when ready. This open-after-its-built approach, an incomplete participation in open source science, generates group-specific *ad hoc* confetti instead of sustainable and integrated platforms for science. We encourage NASA science teams to responsibly contribute back to open source by engaging with the allied communities before, during, and after building technology solutions. NASA should **fund efforts that commit to making upstream contributions and building the minimal amount of project-specific technology possible.**

2023, the *Year of Open Science*, is also the 75th anniversary of the transistor[10]. NASA's open source science initiative extends the USA's mid-century science leadership[11] and heralds a new era of ever-improving tomorrows built by people working together using technology and social innovations.

Citations

[1] *OSTP's Arati Prabhakar Speech at AAAS*, (Oct. 21, 2022). Accessed: Feb. 15, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=0hq8zlSGZqg

[2] "FACT SHEET: Biden-Harris Administration Announces New Actions to Advance Open and Equitable Research | OSTP," *The White House*, Jan. 11, 2023. https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/ (accessed Feb. 03, 2023).

[3] *Scientific Collaboration & Partnership: NASA 3rd Eddy Symposium*, (Aug. 16, 2022). Accessed: Feb. 15, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=izLAVVP6ji4

[4] *3rd Eddy Cross Disciplinary Symposium Wrap Up 2022*, (Jul. 07, 2022). Accessed: Feb. 15, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=8wKtC70LthI

[5] 2i2c, The Carpentries, CSCCE, Invest in Open Infrastructure, MetaDocencia, and Open Life Science, "A Collaborative Interactive Computing Service Model for Global Communities," Aug. 2022, doi: 10.5281/zenodo.7025288.

[6] Executable Books Community, "Jupyter Book." Zenodo, Feb. 12, 2020. doi: 10.5281/zenodo.4539666.

[7] C. L. Gentemann *et al.*, "Science Storms the Cloud," *AGU Adv.*, vol. 2, no. 2, p. e2020AV000354, 2021, doi: 10.1029/2020AV000354.

[8] F. P. Brooks., "No Silver Bullet Essence and Accidents of Software Engineering," *Computer*, vol. 20, no. 4, pp. 10–19, Apr. 1987, doi: 10.1109/MC.1987.1663532.

[9] V. E. Ledvina *et al.*, "How open data and interdisciplinary collaboration improve our understanding of space weather: A risk and resiliency perspective," *Front. Astron. Space Sci.*, vol. 9, 2022, Accessed: Feb. 15, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fspas.2022.1067571

[10] M. Riordan and L. Hoddeson, *Crystal fire: the birth of the information age*, 1st ed. New York: Norton, 1997.

[11] V. Bush, "Science the Endless Frontier," Jul. 25, 1945. https://www.nsf.gov/od/lpa/nsf50/vbush1945.htm (accessed May 24, 2021).