***This is a modified version of the original grant meant for public sharing. It retains the grant narrative while removing the specifics around budget, metadata, reviews/comments, etc.***

# A Collaborative Interactive Computing Service Model for Global Communities

A collaboration between: *2i2c*, *The Carpentries*, *The Center for Scientific Collaboration and Community Engagement (CSCCE)*, *Invest in Open Infrastructure (IOI)*, *MetaDocencia*, *and Open Life Science (OLS).*

## Grant narrative

### Background and context

#### Open science is growing in the cloud

Like many other scientific fields, the biomedicine community is undergoing a rapid expansion in the scope and size of data that are used to answer scientific questions. At the same time, the scientific community feels a new sense of urgency to collaborate across institutional and regional boundaries, to share their work openly and reproducibly, and to share their workflows in a way that others can learn from and build upon. In order to do so, the biomedical community needs improvements in both technical infrastructure and human systems and services around that infrastructure in order to accelerate discovery and enhance the dissemination of knowledge.

In the past several years, many scientific fields have seen the emergence of *Open Science Cloud Services* (OSCS) catalyze and enhance many of these improvements. Open Science Cloud Services allow communities to benefit from the flexibility and scalability of the cloud while adhering to Open Science principles such as community leadership, vendor-agnosticity, and open-source infrastructure[1]. While most cloud services build on proprietary and vendor- or cloud-specific tools with significant risks in vendor lock-in and sustainability, OSCS are managed cloud services with a commitment to using infrastructure and practices that align with open science principles. They accelerate scientific discovery, foster open collaboration for workflows for interactive computing[2], and act as a mechanism to broaden access to computational environments, interfaces, data, and computing power[3]. Examples of successful OSCS platforms are the Pangeo project for large-scale geospatial analytics, the Berkeley DataHub for introductory data science learning, and the Syzygy Project for nation-wide access to cloud computing in Canada.

---

[1] https://openscholarlyinfrastructure.org/
[2] https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020AV000354
[3] https://www.amacad.org/news/international-research-emerging-science-partnerships and https://www.unesco.org/en/articles/unesco-launches-global-call-best-practices-open-science

Recently, several leaders of these projects founded 2i2c, a non-profit organization that develops and manages OSCS for interactive computing. It hosts community-driven computing environments with key open science infrastructure such as the Jupyter Project and the Scientific Python ecosystem via a service model that follows Open Science principles and gives communities the Right to Replicate their infrastructure. This allows communities to benefit from cloud infrastructure by partnering with a non-profit that follows Open Science principles.

## Cloud-based workflows are growing in biomedicine

Cloud infrastructure is becoming increasingly important for modern biomedical workflows. Whilst a "traditional" view of a biomedical researcher might involve pipettes and lab coats, the sheer volume of data produced in modern methods would be impossible to analyze without computational tooling. For example, as of August 2022, the European COVID-19 data portal has over 12 million viral sequences and 4 million biological samples, and the NCBI Gene Portal lists over 38 million genes across 36 thousand taxa.

In other cases, high-throughput cloud computing allows jobs to be run on multiple machines at once, or on particularly high powered computational infrastructure. A job that might have taken days or weeks on a desktop can suddenly be completed in a matter of hours in the cloud. This difference is often big enough to perform analyses that were once considered "unfeasible". An example of this is the 9.5 million hours of computational genomic analysis in Drosophila, Influenza, and Homo sapiens, run over three years. This is equivalent to a computational analysis running for over *1000 real-time years* without high-throughput techniques!

Additionally, cloud infrastructure provides a shared space where biomedical researchers can share workflows and collaborate more effectively with one another. This accelerates learning and sharing of knowledge, which encourages shared understanding within fields and encourages collaboration. For example, the NeuroHackademy has been running collaborative learning and collaboration workshops each year, where most computational work is done via a JupyterHub hosted in the cloud with connections to large datasets[4].

Finally, the benefits of shared cloud infrastructure are particularly useful for historically under-served communities in biomedicine, such as those included in this proposal. These communities often have extra challenges in accessing the tools and data needed to leverage these modern workflows. For example, they may lack the compute capacity, internet connectivity, and/or relevant skills and experience to adopt data-intensive research methods. OSCS infrastructure can make these workflows significantly more accessible and robust by utilizing the fluid and flexible manner in which it makes infrastructure available at a global scale. For example, some long-running analyses will fail critically if the power is lost while running on local infrastructure. If instead the data is already online in a stable datacenter accessible via cloud

[4] See this paper about the hackweek model that Neurohackademy champions, and this pre-print on cloud computing in neuroscience that describes how the cloud can enable collaborative science in the neuroscience community.

infrastructure, then complex and long-running calculations may continue running even in the event of local power outages or internet loss.

## Bringing the cloud to a global community

In November 2020, 2i2c received seed funding from CZI to grow its capacity to develop this OSCS delivery model. 2i2c has spent the last year running pilot projects with universities, research groups, and educational institutions with the goal of developing its infrastructure, service, and sustainability model. It currently serves cloud infrastructure hubs for over 35 communities in the research and education space. It has also gained experience serving the biomedical community via running an OSCS for NeuroHackademy, a two-week distributed learning and collaboration workshop for the neuroinformatics community[5].

Thus far, the communities that 2i2c has served are *largely restricted to US, Canada, and Western European countries*. 2i2c wishes to serve a **global audience** with its OSCS model, but it has identified several barriers to doing so. These barriers include:

- **Cost of running infrastructure**. 2i2c's compensation strategy aims to pay competitive salaries in the technology industry and follows equal pay for equal work, meaning that it pays salaries that follow standards in cities like San Francisco or New York. This results in service costs that are too high for communities with fewer financial resources.
- **Technical shortcomings.** The Jupyter and open source ecosystems enable many new workflows to boost the collaboration and productivity of research and education. However, there are still technical barriers to more effective workflows with community-centric infrastructure. Overcoming these barriers can be partially addressed with new development.
- **Lack of teaching and learning material for OSS workflows in the cloud.** The diversity and complexity of open source tools and workflows can be intimidating and confusing to communities that are not familiar with Jupyter, cloud infrastructure, and open workflows. In addition to simply getting *access* to cloud infrastructure, they would also benefit from guidance, educational materials, and training materials.
- **No model for participatory cloud services.** While there is a rough idea of the skills needed to run cloud infrastructure for other communities, there is not a defined model that provides explicit points for participation and collaboration from multiple stakeholders (such as the one proposed in this grant). Defining this structure will make it easier for communities to understand how they can and should collaborate in providing this service, and will make it more sustainable to maintain and scale.
- **Bridges to communities and their leadership**. 2i2c has limited connections to community leaders within communities outside North America and Western Europe. While

[5] From Ariel Rokem, the lead PI on the NeuroHackademy: "The NeuroHackademy served about 100 participants, half of which were remote. The hub served about 1TB of shared datasets, many of which were staged **as the event was taking place**. About a dozen different instructors pushed code to a single "curriculum" repository that included notebooks on a variety of topics. The hub image was continuously updated with software dependencies in this repository." Moreover, "We had 20 applications from Latin American countries, which is more than 10% of the applications."

a globally distributed team of engineers can develop and manage the OSCS, the leadership for communities that use this infrastructure should come from within those communities themselves. Community leaders are best-positioned to ensure that training materials and content is adapted to the needs, language, and collaborative styles of their community.

Each of these barriers are amplified for under-resourced communities, such as non-profits, community colleges, or institutions in low- to middle-income countries (e.g., most Latin American countries)[6]. They often do not have the budget to pay service fees that are sustainable for the cloud infrastructure that 2i2c provides, and they often lack training content that is catered to underserved groups (e.g., in native languages and developed considering local computational and access constraints).

In order for the biomedical research community to maximize its impact in understanding and curing disease, *it is crucial to leverage the expertise and skills of the global community.* Moreover, in order to ensure that the knowledge gained through biomedical research is accessible and useful to the global community, it is crucial that our infrastructure for scientific research and collaboration becomes globally inclusive by design. This takes many forms, for example by honoring the principle of *Nothing for Us Without Us*[7] by including learning from the communities we seek to serve starting on day zero (i.e., at the time of writing this proposal).

## Proposed work: Interactive Computing for communities in Latin America and Africa

This proposal aims to accomplish four goals, each spearheaded by a team of collaborators on this grant:

- **Goal 1:** Deploy and manage open cloud infrastructure for under-resourced communities in Latin America and Africa.
- **Goal 2:** Create training and pedagogical content to assist others in using this infrastructure for cloud-based science workflows
- **Goal 3:** Build capacity for technical, pedagogical, and leadership skills within these communities.
- **Goal 4:** Identify a participatory service model to sustain, scale, and generalize impact for global communities.

We will describe how these come together in each major goal below.

[6] See https://www.nature.com/articles/d41586-019-01314-3
[7] See https://en.wikipedia.org/wiki/Nothing_About_Us_Without_Us.

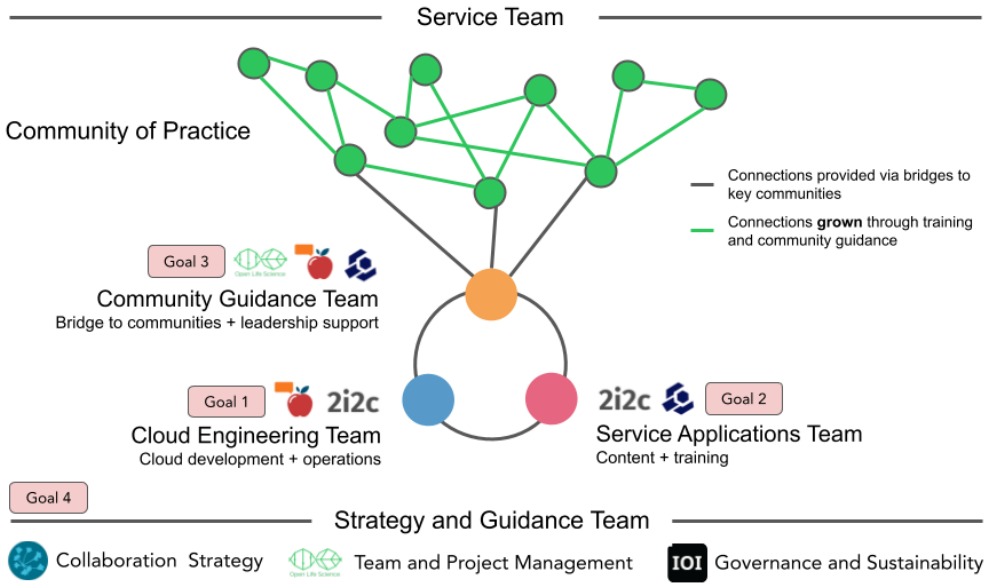


***Figure 1***. Overview of major teams and goals they aim to achieve. The Service Team (upper section) is composed of engineering, service applications, and community guidance expertise. These three areas collaborate with one another in the OSCS. The Community Guidance Team (orange) provides a bridge to communities of practice (top), as well as support and guidance to grow community structure and open practices. The Strategy and Guidance Team (bottom) provides strategic support and design for collaboration and community structure, sustainability and governance practices, and team coordination and management to ensure we achieve our intended impact.

## Community Archetypes: Data-intensive research communities

As part of this effort we aim to define *community archetypes* to guide decisions about what communities to reach out to and how to define impact. This project will focus on **data-intensive research communities in Latin America and Africa.** Moreover, our community engagement efforts will focus on research communities that **intersect with biomedicine**. We describe two examples below.

Example 1. In many bioinformatics fields such as biomedicine, computational research skills are essential to modern research. Learning these skills depends on access to **powerful and specialized compute infrastructure** due to the complexity of software environments and the size of datasets associated with analysis of data from high-throughput methods such as single-cell genomics. This presents an obstacle to early career researchers in under-resourced communities, to whom these compute environments may not be available[8].

Example 2. **Collaborations around large datasets** that are based in the cloud–for example, distributed geospatial analytics communities that work with satellite imagery data. Accessing these datasets in an interactive fashion for data analysis is traditionally very time consuming and requires technical expertise. However, the modern cloud-native data stack builds upon tools and

[8] See https://www.nature.com/articles/d41586-019-01314-3

standards that make this much easier, such as XArray, Zarr, and Dask. These tools, along with managed cloud infrastructure with pre-defined environments and access to datasets, can drastically reduce the barrier to exploring, asking questions with data, and sharing with others. This stack has also recently been explored in the biomedicine community via collaborators on the Neurohackademy, and we are confident it will prove useful for the biomedicine community in general (for example, the genomics community also has a need for high-throughput analysis of very large datasets).

To serve these types of communities, we will offer a combined socio-technical service that includes cloud infrastructure, end-user training material creation, capacity building for training, and community leadership training. The types of cloud infrastructure we serve will be utilized in the following primary workflows:

1. During **Train the Trainers workshops**, community leaders undergo learning in how to use these tools themselves, and how to effectively teach others how to use them.
2. During **End-user training workshops**, community leaders learn skills in cloud workflows in order to share them with others in their community via their own workshops and events.
3. Hosting infrastructure for **ongoing research collaboration** for a small subset of communities as a pilot. While the goal of this grant is focused around training and capacity building, we have found that training events naturally lead to a need for community use *beyond the training period*. We believe this is an opportunity to explore how to further support the work these communities do, and to grow more connections within communities over time.

We describe the technical and social aspects of this service in more detail below.

## Goal 1: Deploy and manage open cloud infrastructure for under-resourced communities in Latin America and Africa.

**Key Team Organizations**: 2i2c, MetaDocencia (via external secondment with the High-Performance Computing Center in Argentina at the National University of Córdoba, Argentina).

### 1.1 Deploy and operate an interactive computing hub for communities in Latin America and Africa

2i2c will provide a dedicated interactive computing cluster that is managed on behalf of communities in Latin America and Africa. Connections to these communities as well as their stewardship and guidance will be led by MetaDocencia (Latin America) and The Carpentries (Africa).

It will follow the principles of OSCS and provide flexible and community-specific interactive computing environments in the cloud. For example, we aim to manage a research-focused JupyterHub for the High-Performance Computing Center in Argentina at the National University of Córdoba (CCAD) in collaboration with Dr. Nicolás Wolovick, its director, and MetaDocencia.

### Goal 1.1 outputs

- Two kubernetes clusters (one for communities in Latin America, another for communities in Africa) that are deployed and managed via 2i2c's deployment infrastructure.
- Community-specific hubs that are deployed on each cluster, and utilized for workshops (and ongoing collaboration in a small subset of cases). Our target goal is 10-15 community hubs per cluster with 1-2 ongoing research hubs in each cluster as a pilot.
- Cloud engineering team provides ongoing support and operational improvements to this infrastructure via a support and operations system.

## 1.2 Improve open source tools that facilitate collaboration and open workflows via interactive computing

As we serve these communities it will naturally create opportunities and ideas for new development and enhancements in order to improve the experience of those using the infrastructure. We especially hope to target new development in areas that enhance collaboration between multiple communities that use the same cloud infrastructure. This collaboration will allow us to perform this work and test out new improvements on the communities we work with.

### Goal 1.2 outputs

- Improvements to the codebase in 2i2c's collaborative infrastructure deployment repository.
- Contributions of code and documentation to open source tools in the Jupyter / JupyterHub ecosystem that are needed to enhance the usage for these communities.

### Goal 1 outcomes

- Key communities have access to cloud-based infrastructure with open source tools for workshops
- Communities become familiar with this infrastructure via using it and incorporating it into their workflow.
- The open source tools under this infrastructure are more effective at supporting networks of multiple related communities.

### Goal 1 indicators

- Workshops have been run that utilize the OSCS hubs that are managed as a part of this project.
- During these workshops, there is a clear increase in utilization of the cloud infrastructure, corresponding to users joining and doing work in the cloud.
- Research and analysis outputs from key communities occur as a result of learning done via the OSCS infrastructure, or in a subset of cases, research conducted via the OSCS infrastructure (papers, figures, blog posts, etc).

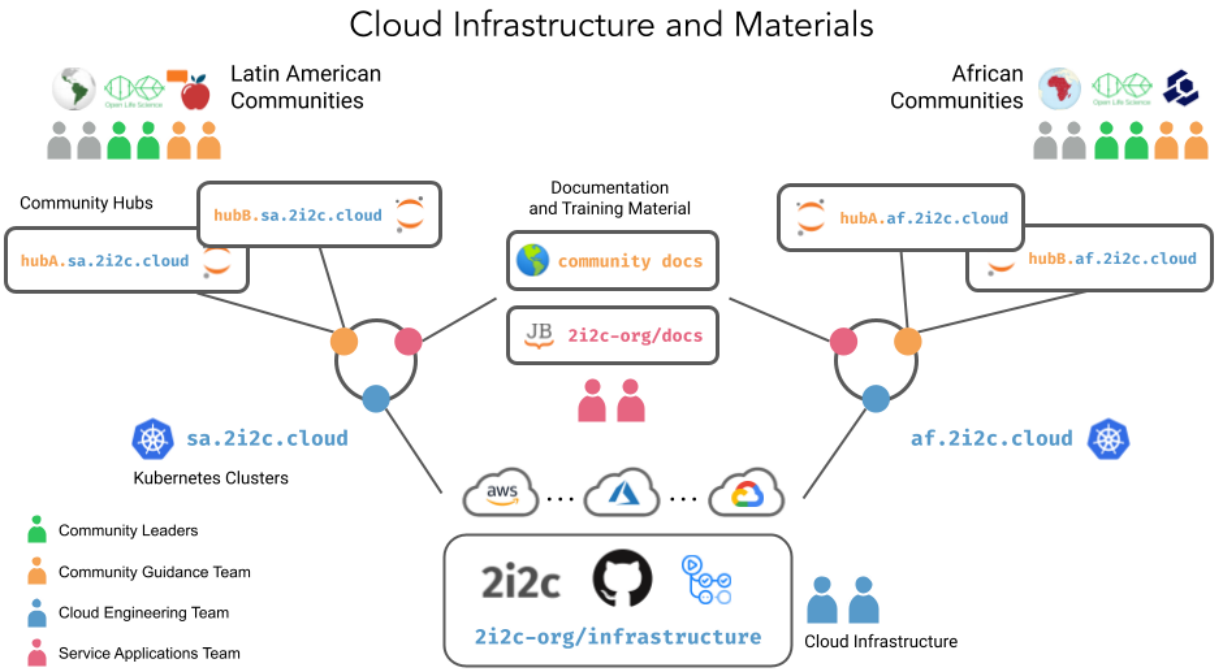


***Figure 2.*** *An example of the major infrastructure and materials that are created and managed. From bottom to top: at the bottom, a GitHub repository managed by the 2i2c Cloud Engineering team contains configuration and deployment infrastructure for several Kubernetes Clusters running on commercial cloud providers. These clusters are dedicated to a community of practice and can contain multiple hubs for sub-communities. A collection of documentation and learning resources (middle) is used and developed by Community Leaders (MetaDocencia / Carpentries / OLS) and Service Applications Specialists (2i2c) to train and guide others.*

## Goal 2: Create training and pedagogical content to assist others in using this infrastructure for cloud-based science workflows

**Key Team Organizations**: 2i2c, The Carpentries, MetaDocencia, OLS

In addition to providing *access* to the cloud infrastructure, it is also crucial that we provide *guidance* around how to use it effectively for open science in the cloud. There are hundreds of tools in the open source ecosystem that might be used via a JupyterHub like the ones we deploy, and communities will benefit from a guided pathway through this ecosystem to achieve specific outcomes. This will reduce the difficulty of *using* cloud infrastructure by creating training materials and conducting training sessions (see goal 3 for more information on training sessions) that guide communities in the effective use of the hubs provided by 2i2c (2.1) and contextualizing the material to best suit the needs of Latin American and African communities for which the hubs are designed (2.2).

### 2.1 Create learning material to guide users in open workflows with cloud infrastructure

We will create a collection of documentation that is meant to guide users in utilizing our cloud-based infrastructure for open science in the cloud. For example, we wish to consolidate and create material around scalable cloud workflows with Dask Gateway and Zarr, teaching models for data science with JupyterHub and Jupyter Book, and collaborative events such as Hackweeks. Building this material will require interfacing with community members in The Carpentries and MetaDocencia, to understand what content is best-created by 2i2c and how it could interface with the content created by other communities.

#### Goal 2.1 outputs

- Web-based documentation for using the cloud infrastructure. This will likely be a combination of tutorials, how-to guides, reference materials, and explanatory documentation[9].
- Workshop-focused materials that are adapted from the usage documentation. These materials will be open, reusable, accessible, and findable including workshop slides, pedagogical narratives, learner personas, organizational templates (e.g., invitation, registration, governance, community paths), and impact measurement instruments (e.g., pre/post workshop surveys).
- Workshops are held which utilize these materials for teaching end-users of cloud infrastructure.

### 2.2 Contextualize this learning material for workshops for communities in Latin America and Africa

In addition to creating base training materials around our managed infrastructure, it is important to *contextualize* this source material for the perspectives of communities that will interact with it. This means providing extra information or adapting source material to consider the context of a specific end-user community. We will prepare our source material in a way that can be contextualized to different global audiences, and will leverage our partnership with MetaDocencia to ensure that equivalent Spanish-language content can be utilized by the Spanish-speaking communities that we serve as a part of this pilot.

In addition, researchers in underserved areas often need to level-up their knowledge **before** learning the state-of-the-art tools. As such, we will create “on-boarding” content that is language-localized to assist others with the early steps needed to make the most use of this infrastructure. Creating this content will also enrich the source material, making it more accessible to a wider audience and providing an opportunity to learn how underserved communities optimize their use of minimal resources. Leveraging the existing relationships that MetaDocencia and The Carpentries have with groups in Africa and Latin America, we will involve communities from underserved geographies early on in the development of material in **2.1** and then in contextualization to increase their participation in the process. This will also maximize 2i2c’s learnings for serving these communities and aid it in building a more global open infrastructure community by the end of this project.

[9] See diataxis.fr for inspiration of this structure.

### Goal 2.2 outputs

- Workshop material is translated into native languages of the key communities in this grant (at least Spanish, with a target of French as well if we can find the expertise in language translation).
- Workshop material is refined and modified as-needed to map onto local community context (e.g., changing examples to fit a local context, or assumptions about access to reliable internet).
- Improvements to the source material in order to make it more accessible and useful for a wider audience.

### Goal 2 outcomes

- Users of OSCS infrastructure provided by 2i2c have a more extensive knowledge-base to guide their efforts in using the infrastructure.
- Users in Latin America and Africa utilize more language-localized content that helps guide their use of the infrastructure.
- Researchers in Latin America and Africa achieve an increased understanding of how to use open source tools and the cloud to do their work more collaboratively and effectively.
- Leaders in 2i2c achieve an increased understanding of how to provide open infrastructure effectively to underserved communities.

### Goal 2 indicators

- Size and scope of documentation that covers common use-cases that are relevant to scientific research in the cloud
- Web traffic indicators to original documentation documentation in local languages
- Community leader attendance at training workshops
- Self assessments and interviews conducted after training workshops

## Goal 3: Build capacity for technical, pedagogical, leadership, and community building skills for global communities.

**Key Team Organizations**: MetaDocencia, The Carpentries, OLS

In addition to building content for end-users of this infrastructure, we also hope to grow capacity for these communities to replicate this kind of service in the future, or to jointly participate in 2i2c's collaborative cloud services in the future. This involves a **knowledge transfer** of skills related to pedagogy (3.1), cloud infrastructure (3.2), and community leadership (3.3).

### 3.1 Build pedagogical capacity and skills within these communities via Train the Trainers workshops

We will create and offer workshops for community leaders that are inspired by The Carpentries' "Train the Trainer" style of community-empowering pedagogy. We will follow a model that The Carpentries recommends, whereby community leaders begin by taking a "Train the Trainer" workshop, then take part in a workshop from an end-user perspective to understand the content

better, and then begin running their own end-user workshops. Train the Trainer workshops will be done via adapting and enhancing materials that The Carpentries and MetaDocencia have created in their experience offering Train the Trainer workshops over many years[10]. End-user workshops will be delivered by 2i2c and focused around cloud workflows created in partnership with 2i2c (Goal 2.1). End-user workshops run by community leaders will then follow, though these will be managed by community leaders and may fall outside of the window of this project.

#### Goal 3.1 outputs

- The content creator for workshop materials will take the Lesson Development training developed by The Carpentries. This training teaches participants methods to develop effective curricula.
- Pedagogical materials are created and improved that provide the building blocks needed to teach communities about this infrastructure (templates for slides, tutorials, etc).
- Community leaders take part in “Train the Trainers” workshops that are run via our team.
- Community leaders also form the first cohorts of *learners* for the use-case tutorials developed in Goal 2.1, thus gaining direct experience of how the tutorials are intended to be delivered before they go on to teach the skills to members of their own communities.

### 3.2 Grow and connect to local expertise in cloud infrastructure management

A core part of 2i2c’s mission is to run the infrastructure that other communities may replicate on their own. As 2i2c runs the cloud infrastructure for these communities, we also aim to **share our knowledge** of cloud infrastructure by inviting participation in our cloud service team to the communities we serve. We will use this project to *learn and pilot* how to do so effectively, and in order to understand any unique considerations to running cloud infrastructure for communities in Latin America and Africa[11].

We will partner with Dr. Nicolás Wolovick and MetaDocencia to explore and create material for how institutions in Latin America can utilize cloud infrastructure to provide similar services like the ones that 2i2c provides. This collaboration will take the form of a *secondment*[12], whereby a member of the CCAD team spends time working with 2i2c’s cloud infrastructure team to understand its operational and infrastructure model.

This content might be in the form of 2i2c-specific documentation, or via upstream contributions to open source projects like JupyterHub. Dr. Angelique Trusler will work with Carpentries member organizations to identify partners in Africa to develop local cloud expertise.

Our goal will be to have a better understanding of the challenges and opportunities for teaching communities about how to provide cloud infrastructure for interactive computing, and to make

---

[10] For example, this post on The Carpentries blog describes MetaDocencia’s experience in teaching Train the Trainers models in Spanish.
[11] For example, this Nature article describes many challenges to running technical infrastructure in Latin America, and has several suggestions and ideas from our collaborator Dr. Wolovick.
[12] A secondment is a process by which an organization’s employee temporarily spends time working with another organization.

some early progress and experiments in doing so. For example, here are a few questions we aim to answer with this pilot:

- How much mentorship and guidance will cloud secondments need?
- How can we balance the learning needs of cloud secondments with the needs of reliability and stability of the communities that 2i2c serves?
- What are the learning goals for the secondment? What do we want the person to achieve by the end of the project?
- What structured material could we create to help secondments make rapid progress within 2i2c and share back to their primary organization what they have learned?
- Should this type of training focus on Kubernetes-based infrastructure like the ones 2i2c manages, or focus on JupyterHub more generally without assuming Kubernetes infrastructure?

#### Goal 3.2 outputs

- A computer scientist from CCAD will work with the Cloud Infrastructure team as a secondment, learning and assisting with running cloud infrastructure to learn how the infrastructure and workflows supported by this service can be translated back to the CCAD context.
- An improved understanding of the ways in which 2i2c's cloud infrastructure could be generalized and adapted to a context in Latin America or Africa.
- Improvements to documentation will instruct others in how to manage the infrastructure in this project (either in 2i2c's team documentation or in community-contributed documentation)
- Particular challenges discovered in serving these communities will be documented and shared broadly via blog posts, website documentation, etc.

### 3.3 Grow open practices and leadership within communities of practice.

Ultimately OSCSs are designed to enable collaborative, inclusive, open communities of practice that leverage shared infrastructure to work with one another more effectively. For computational research projects to be open, collaborative and sustainable, an understanding of and experience with open and inclusive community practices are as crucial as having the appropriate technical infrastructure.

We will partner with the existing open science community leadership training organization, OLS, to help those who join this infrastructure program to develop healthy open community and leadership infrastructure that supports them in sustaining volunteer contribution and engagement beyond this program.

Through five cohorts, OLS has trained 252 participants from 6 continents, across 47 low-, middle-, and high-income countries, and has significant working experience in Latin America and Africa, collaborating with The Carpentries and MetaDocencia.

OLS will apply localisation and contextualisation techniques to deploy a (Spanish/French-specific) deep-dive condensed version of their leadership training, and invite participants to join the global 16-week open science community training/mentoring for additional networking opportunities. Participants will be mentored in their preferred language with each session facilitated by open science practitioners to provide:

- Expert talks and demonstration of open community-building models, open source tools, for open science techniques from project leaders from the Global South
- Introduction to open source tools and communities from biomedical science
- Structured group discussion to connect with other participants, explore and strengthen the understanding of how open science applies to their projects, and foster opportunities for further collaborations
- Assignments and frameworks to guide project development, maintenance and sustainability plans
- Integration in the global OLS community, to further knowledge exchange, collaboration, and mutual support
- Further information about the program: OSCS Proposal - OLS expanded role

### Goal 3.3 outputs

- CC-BY licensed, publicly accessible and reusable open source science community leadership training materials (slides, worksheets, resources, further reading) designed by and for Latin American and African researchers, with cultural and technical awareness that doesn't default to Global North norms, and is in contextually appropriate languages, e.g. Spanish, French, Portuguese.
- Openly shared high-quality recorded training videos of expert speakers sharing short talks (~15 minutes) on these topics, with captions in relevant languages, translated into English when needed.
- A short report from each participant at the end of the programme to be shared under the CC-BY License. They will receive certificates and complete their training by presenting their learning in a graduation call (video recorded and shared).
- One 2-week deep-dive online open science leadership training camp with Spanish-language-specific mentoring and talks, additional 2-week camps in French if further funding for staff time and preparation is available. All participants will be invited to co-deliver this camp and share their learnings with new learners.
- Local meet-ups or events organized by graduates in their network to share their lessons learned. They will also be invited to contribute to the original OLS training materials and supported to disseminate these resources in their communities.

### Goal 3 outcomes

- Technical leaders within these communities have an increased understanding of interactive computing services delivered via cloud infrastructure.
- Leaders within these communities are empowered with the knowledge and pedagogical skills to teach others in their communities.

- Leaders within these communities have an enriched understanding of how to grow open and inclusive communities.
- Communities that participate in this project have more, and stronger, connections with others, and an increased culture of collaboration and mutual support.

### Goal 3 indicators

- Extent and depth of open science and technical leadership knowledge in interactive computing services, including in its use and its delivery in their local context. This will be quantified and assessed by means of including an end-of-training survey, a short format report by participants, as well as their subsequent participation and leadership roles in the community in the long term (indicated by the number of participants and their projects).
- Community training workshops that have been run by community leaders, without direct management or instruction from this team.
- Number of participants returning to co-facilitate the open science training camp and related workshops, new learners joining the training camp and community leaders promoting the content (counted by citation and use via Zenodo, as well as targeted survey)
- Opportunities for collaborations facilitated between the participating communities. This will be indicated by the number of collaborative efforts in delivering training workshops that apply the learned pedagogical skills, as well as the size of the user base for the cloud infrastructure.
- Number of members and contributors in the resulting open and inclusive communities and projects, ensuring their long-term maintenance and sustainability. This can be planned as a long term impact assessment process generally designed for this project.


***Figure 3.*** *An overview of the major roles involved in serving a single community via an Open Science Cloud Service. Each role is described with a color and a person symbol. Roles contain a*

*one sentence description of the type of person that excels in this role. Between each role are a collection of collaborative actions that two roles engage in to serve the community according to their skills and abilities.*

## Goal 4: Identify a participatory service model to sustain, scale, and generalize impact for global communities.

**Key Team Organizations**: 2i2c, CSCCE, Invest in Open Infrastructure, OLS

Our ultimate goal is to identify a model for running OSCS for global communities that is *generalizable, sustainable, and replicable*. This project will allow us to gain real-world experience in serving these communities, and to experiment with various ways to collaborate with and serve them. It is equally important that we identify a model for continuing the impact of this project beyond the scope of the grant window. We believe that this will be useful to a broad community of stakeholders that are interested in serving global communities with open infrastructure.

To ensure that we learn from this experience, and that our learnings are shared with others, we must engage in strategic design and coordination for the service, and share the design and coordination outputs and practices for broader reuse. Wherever safe and appropriate, we will conduct our own work in the open and publish our team strategy, operations, and learning in an accessible format (for example, similar to how 2i2c manages a Team Compass as a public record of its policies and strategy).

To accomplish this we will partner with three leading organizations to provide strategic support and guidance. CSCCE for community structure and management (4.1), IOI for community governance and sustainability (4.2), and OLS to oversee our project operations and ensure we document our learnings in a way that is accessible to others (4.3). IOI and CSCCE will jointly support the project manager to ensure that they can effectively implement the community, operational and governance recommendations as described in Goals 4.1 and 4.2 below.

### 4.1 Identify and prototype collaboration structures that connect the infrastructure team, service applications team, and community leaders.

Successfully running these services will require developing a collaborative service model that shares responsibility between 2i2c and the communities it serves. It will require designing roles, team structures, and expectations that allow each group to utilize their strengths, and that centers the community as the leaders of the service. In order to do this, we will need to coordinate our external facing work as well as align internally as a project team to be able to deliver the different work streams.

In order to facilitate this, we will partner with CSCCE who will run a series of project kickoff meetings to support the alignment of the project team members around shared language, norms of engagement and processes, which will aid delivery of the work. CSCCE will then shift to providing mentorship to the newly-hired project manager and community coordinator as well as advice on overall community design, information flow, and content management. This mentorship

will take the form of supporting the community coordinator in designing scaffolding material templates for local communities and helping them to lead activities where the local leaders contribute to the creation and maintenance of this documentation. It will also include guidance on supporting the reuse of technical materials that community leaders may create during the project e.g., by synthesizing, curating and templating materials as they emerge.

#### Goal 4.1 outputs

- CSCCE leads a series of kickoff meetings to align project team members around shared vocabulary, norms of engagement, and project management processes - which will result in documentation to support the project (codified in a team playbook).
- Sections of the team playbook e.g., a glossary of shared terms, the core values underpinning the project, and communication channels may be shared publicly. Blog posts describing the process of alignment will be used to highlight the resources and how to use or replicate them.
- The OLS project manager and community coordinator will be a CSCCE-certified community manager - gaining specific deep skills in communications strategy, online community management and team coordination.
- Clear programming plans for the community coordinator so they can effectively guide the community leaders e.g., an onboarding pathway that connects them to OLS, The Carpentries, and MetaDocencia activities in a coherent manner.
- Scaffolding documentation for guiding and supporting community leaders, created via strategy development with the community coordinator.

### 4.2 Identify and prototype multi-stakeholder governance and sustainability structures for participatory cloud services.

In addition to collaboration structure, it is also crucial this project be structured to adequately and meaningfully consider governance and accountability in running participatory services, particularly as they pertain to communities that are historically excluded or under-represented in many global infrastructure projects. To ensure the work done on this project has enduring value, the project also needs to critically understand the necessary preconditions for sustainability and design the project to be both highly collaborative and viable in the long-term.

To address these questions, we will partner with Invest in Open Infrastructure to provide research, recommendations, support, and ongoing evaluation to implement best practices for effective and inclusive governance of the project, as well as meaningful sustainability to realize long-term value from the work done as part of this project. Below are some questions that we will explore in this work:

- What kinds of governing structures strike a balance between the many different communities that use the infrastructure, but still give them representation?
- What governance models best support the intended level of community engagement and stakeholder participation?
- How can we intentionally balance power dynamics between communities that may have large differences in the resources available to them? Particularly, what are governance

practices that can best-ensure that the needs of traditionally excluded/marginalized communities are represented and served?
- How can the project teams and partners achieve a meaningful level of operational transparency that empowers stakeholders to engage with the operational and strategic planning for this project?
- What resources are necessary to effectively resource the inclusive and sustainable governance of a diverse, multistakeholder collaborative project for inclusion and long-term viability?
- What governance process should we develop to select tools and services that will ensure alignment of these selections with our values of openness and inclusivity while delivering value to our project partners?
- What ongoing challenges of governance and management can and should be addressed by further research, experimentation, and iteration?

IOI and project leadership will conduct initial research in governance activities similar to the one being undertaken in this proposal. Based on those findings, IOI will design an early-stage governance and sustainability plan, which will be piloted by project teams and participating community members. IOI will participate in ongoing consultation with the project teams, documenting learnings and assessing the efficacy of the initial plan as the project develops in order to test and validate interventions to improve the governance and sustainability of the project. At the conclusion of the project, IOI will provide a summary of findings documenting the initial plan, the lessons identified in the course of the project, and a final set of recommendations for further reflection and development.

Through co-designing, refining, and iteratively developing this governance and sustainability model with the participating communities, we not only motivate participating teams and individuals to reflect more deeply about governance and power dynamics within their own communities and networks and maximize their learnings, but also build co-ownership into the resulting model and encourage its reuse and further generalization in the long-term by other projects, advancing the state of knowledge and experience with more robust, inclusive, and engaging governance structures that will provide the necessary ingredients for long term viability.

### Goal 4.2 outputs

- IOI will draft and release a summary of relevant research and initial recommendations for architecting the governance and sustainability for the project
- IOI will draft and release an initial governance and sustainability plan for the project based on the findings of the research summary, with clear implementation guidelines so that the project manager and coordinator can effectively execute the plan
- IOI will participate in meetings with the project teams to gather qualitative and quantitative feedback formally and informally to inform proposed changes to the governance and sustainability plan, to be socialized with the project teams
- IOI will provide a final summary report of governance activities, including documentation of the initial plan, changes to the plan over the course of the project, and final recommendations based on the findings from working with the project teams.

### 4.3 Generalize these experiences to a sustainable model for providing participatory cloud services to global communities

Finally, we believe that it is important that our team embody the same principles of open practices that we wish to cultivate in other communities. This will allow us to make our experience for participatory open science cloud available to the world to learn from, in the hopes that it may be replicated and expanded to other countries, workflows, research domains, etc. To accomplish this, OLS will provide Project Coordination to ensure that our outputs are maximally accessible and useful for our team and for others. They will receive strategic support and training from CSCCE and IOI to empower them in this role.

#### Goal 4.3 outputs

- Public CC-BY licensed documentation about the strategy, operations, and inner-workings of our service model (potentially in the form of a “Team Compass”)
- Synthesis documents that summarize and describe the major learnings of this project, and our proposed next steps moving forward (this may take the form of a research paper or published work, we will choose the appropriate output to maximize the impact of our learning).

### Goal 4 outcomes

- The teams participating in this project work more smoothly together and have a well-aligned, trust-based team that is able to work together successfully on the delivery of the project, including with effective asynchronous communications and project management processes.
- Understanding of how we work, what we are trying to achieve, the major challenges with serving these communities, and a strategy for how to overcome them.
- A shared language for talking about the project and its component parts, including clarity around the use of words such as community, member, engagement, sustainability and governance.
- Team members and community leaders in this project appreciate the importance of effective and meaningful governance and an open, transparent way of working, and are inspired to adopt what they have learnt and created in this project in other communities and collaborations that they take part in.
- There is a community and service governance model that effectively empowers and provides representation and accountability to the communities we serve.
- The service that we describe in this project is more sustainable, scalable, and generalizable to new communities.
- Other organizations and communities are inspired and informed by the experience of this team in order to create their own open science cloud services.
- The community coordinator will be equipped to empower community leaders to lead their communities via the creation of clear onboarding pathways that connect the leaders to the other project partners.

- The community coordinator will be empowered to work with community leaders to create reusable scaffolding documentation to support their community work.

### Goal 4 indicators

- Results of a team culture survey demonstrate that members of the project team feel included, aligned and happy with their roles on the project and their interactions with others.
- Communication about the project, including external communications such as blog posts and conference talks will use a common vocabulary, making the goals and activities of the project clear to others.
- More explicit conversations across the ecosystem about how intentional project design, team orientation and coordination contributes to more inclusive and effective teams.
- Effective empowerment of community leaders - with minimal confusion about project partners, available resources and how to communicate about the project to their own communities.
- Minimal burnout and/or inefficient use of volunteer time because appropriate scaffolding supports volunteer focus on the things they care about - not recreating shared materials and processes.
- The number of organizations/service providers expressing interest in reusing and adopting the output service and operational models based on interviews and survey responses
- The number of organizations/service providers creating their own open science cloud services, inspired by 2i2c's model as indicated in interviews, survey responses, or other verifiable feedback,
- Individuals, teams and external organizations that report deriving inspiration from this collaboration's strategy or operations
- Overall engagement and satisfaction with the project's governance structure, as indicated by the team culture survey, qualitative feedback regarding team interaction and ways of working

## Value to biomedical communities: A service model for global biomedical research communities and beyond

Together, we believe that these four goals will complement one another to provide a flexible, scalable, and generalizable socio-technical model for cloud infrastructure services that can be used across many biomedical communities in Latin America and Africa, as well as in the global community. They do so by bringing together four things:

- **Access** (Goal 1) to cloud-based infrastructure and workflows through hosted infrastructure. This will lower the barrier to accessing and utilizing scalable cloud infrastructure for computational methods that require scalable computing or access to large datasets. For example, by providing a scalable Dask cluster to a genomics researcher in Argentina that must analyze many GB of data quickly.

- **Training** (Goal 2) to shorten the learning pathway to utilizing this cloud infrastructure for use-cases in biomedicine and beyond. For example, by providing this researcher a set of Spanish-language documentation for scalable genomics computation using Dask in the cloud, which they can use alongside their infrastructure from Goal 1.
- **Community** (Goal 3) to provide a cultural context and leadership that enables open and collaborative practices via the cloud workflows we teach. For example, by empowering leaders in this researcher's community to connect them with other trainings and researchers with similar interests who are using the same cloud infrastructure.
- **Strategy** (Goal 4) to provide intentional design and structure to this service to ensure it lives up to its principles of representing its stakeholder communities, and to provide a formal description of the service that can be generalized to others. For example, by demonstrating how the cloud service (Goal 1), training (Goal 2), and community structure (Goal 3) around this researcher could be sustained beyond the timeframe of this grant or generalized to adjacent communities.

We believe that each of these individually represents a significant benefit to communities in biomedicine, Latin America, and Africa. Taken together, they will build on top of each other to provide support to researchers at each stage of the learning and research workflow, and to create a technical and social foundation that will increase the chance of ongoing and fruitful collaboration within and between these communities.

The tools, skills, and community created by this project will be an important step forward in helping global communities in biomedical research begin to incorporate cloud-based workflows into their research. It will improve their ability to ask research questions and share their work with others by utilizing a common framework of tools and practices in a growing network of community leaders. It will also help catalyze a growing understanding of how cloud infrastructure fits in with their workflows, and how services and technology can be designed to utilize the cloud in a way that is globally inclusive and equitable. This will lead to biomedical research that is more impactful and that benefits a global community.

# Grantee information and resources

## Grantee qualifications

Below we list the major collaborating organizations in this proposal, as well as their core areas of expertise.

### Qualifications for necessary work

Below we describe the major programmatic areas and competencies of each organization in this collaboration, and why they are poised to have impact on this project.

**2i2c** is a non-profit initiative dedicated to making interactive computing more impactful through community-centered open infrastructure services. It integrates, customizes, and manages cloud

infrastructure that facilitates open and collaborative workflows in research and education. It is also manages this infrastructure in a way that aligns with open practices, emphasis community-driven technology, and supports open source communities that create this infrastructure. 2i2c manages cloud infrastructure for over 35 research and education communities, from collaborative workshops like NeuroHackademy, to individual research teams like the MEOM satellite imagery team at Grenoble, to university-wide education such as the University of Toronto, to distributed research communities like Pangeo project. They are also core contributors of the Jupyter project, with a focus on cloud infrastructure and the JupyterHub and Binder ecosystems.

**MetaDocencia** is a collaborative and inclusive community built to support Spanish-speaking educators. MetaDocencia (which in English means “meta teaching”) was born in the wake of the COVID-19 pandemic as a response to the need and urgency for incorporating tools for a quick and effective switch to online events, understanding that contexts of isolation increased limitations in accessibility and inclusion both for students and trainers. We share evidence-based teaching methods and develop open, reusable, and accessible resources to foster effective teaching practices, useful for online and traditional classrooms alike.

During our first two years we developed 5 short, open courses that were offered in 81 free editions, mostly in Spanish, reaching 1,163 educators from 30 countries. Our hallmark course is the Introduction to Online Teaching Essentials, a 3-hour workshop designed to introduce basic good practices for management of online classes, meetings, and other events. Another successful workshop is How to Teach Programming Online, which covers techniques and good practices to design and implement programming courses, including useful tips to get feedback and evaluate students. Those who took our courses showed a high level of satisfaction (Net Promoter Score > 80%) and found the contents practical, useful, and novel (97% shared that they learnt something new). We also completed 3 Train the Trainer events in Spanish in partnership with The Carpentries. This helped more than 50 new people to be certified as trainers in teaching programming and data science skills to researchers. We are currently conducting an 8-part open webinar series to discuss how to build governance of Latin American communities of practice. Throughout all these activities, we also worked towards improving the accessibility of all our events and supporting materials, and to document our knowledge and experience openly for the benefit of other researchers and Open Science communities.

MetaDocencia has been mostly based in Argentina but quickly expanded to Latin America and the Caribbean, with presence even in regions where Spanish is not the main language, like North America, Europe, and Australia. Maintaining an active collaboration with communities such as R-Ladies and Women in Bioinformatics and Data Science Latin America has been instrumental for this growth, as well as the support from Code For Science & Society and CZI. We maintain an open Slack workspace with over 650 participants that serves as our main interaction hub, and we engage with more than 2,500 followers through our social media and YouTube accounts. These people are generally educators, researchers, and professionals in STEM disciplines and others such as Social Sciences, Humanities, Economics, and Arts. This diversity of backgrounds and interests encourages a cross-disciplinary knowledge-sharing on educational and technical practices. For example, we offer extended support on Slack to people who are willing to learn

more about a certain programming resource or need assistance when implementing what they experienced in our events.

**Open Life Science** (OLS) currently offers a 16-week training and mentoring program for researchers and academic leaders to create, lead, and sustain open research projects and become open science ambassadors in their communities. Through five cohorts, OLS has trained 252 participants from 6 continents, across 47 low-, middle-, and high-income countries.

OLS graduates are invited back into leadership roles in OLS to strengthen their knowledge in open science and extend their network by re-joining subsequent cohorts as mentors, call facilitators, and expert speakers. Previous OLS graduates have gone on to be awarded grants from European Open Science Cloud (EOSC), Code for Science and Society, Lacuna Fund, the Awesome Foundation, and fellowships from the Software Sustainability Institute and Open Bioinformatics Foundation.

With the combination of practical training by experts and the conceptual knowledge of OLS mentors, participants are guided to **reflect on and apply open practices in the context of the socio-technical environment** where they conduct their research. Many participants not only lead open science projects on tools and resources but create **Communities of Practice around their projects**. This has been enabled by the strong emphasis on **open community leadership** throughout the OLS curriculum via targeted training and structured assignments. Frameworks provided in OLS as well as contextualized mentorship from expert practitioners allow project leads to design their projects and communities **in the context of a larger open biomedical science ecosystem** while contributing to the local capacity building. Being a part of the OLS network allows enhanced discoverability of opportunities for funding, collaboration and future directions for participants as they establish themselves as open source practitioners.

OLS delivers professionally facilitated training to enable open science knowledge exchange and equitable collaborations between communities while centering marginalized voices and interests. In cohort-based training and mentoring, OLS achieves this by applying a three-prong approach that involves:

1. demonstrating best practices and modeling positive behaviors for enhancing accessibility, inclusion and multi-channel participation in online spaces, such as by providing closed captioning, live-translated bilingual calls, asynchronous interaction platforms and hosting coworking calls.
2. spotlighting a range of different case studies led by diverse community leaders, offering dedicated modules for discussing Equity, Diversity and Inclusion, highlighting the importance of mental health in research, and offering Ally Skills training to use our societal advantages to support others with less privilege.
3. providing targeted financial support such as stipends, microgrants and honoraria to foster equitable engagement from members in marginalized or lower-resource settings for purchasing hardware like headsets and webcams, paying for expensive internet for low-income members, and covering the cost for electricity generators where participants

deal with power cuts, offering the cost of childcare, or to recompense open science-related volunteer work.

This approach has been proven successful by the number of highly diverse communities that stay participating at OLS after each cohort, and the new communities in remote areas that keep coming to new cohorts.

**The Carpentries** has demonstrated that evidence-based teaching enacted through hands-on, interactive workshops is effective in transforming researchers' use of computational best practices for data analysis and visualization, including both short- and long-term changes in learners' behavior and attitudes[13][14]. Their instructional model is based on targeted, real-time support for individual learners. They actively recruit Instructors and Trainers with diverse backgrounds and from a diversity of geographical locations and work with member organizations to reach diverse and historically underserved communities around the world. Since 2012 and as of the end of 2021, The Carpentries have run 3,000 workshops in 64 countries and trained 3,419 volunteer Instructors at their 92 member sites. Through support from another CZI grant (announcement), The Carpentries is providing three memberships to institutions serving underrepresented communities in Central and South America and Southern Africa.

The Carpentries community has designed and developed, using its collaborative infrastructure, ~50 open source lessons delivered in hundreds of workshops annually. They have a community-developed Code of Conduct (CoC) and provide CoC training to all Instructors and Trainers. They train all new Instructors on the technical aspects of lesson development through GitHub so that everyone has the relevant skills to contribute to their lessons. Like any open source project, it is important that there are clear pathways for volunteer engagement. They have active channels for communication, including email lists, GitHub issues, a blog, and a newsletter to maintain active communication to and within the community.

The Carpentries and the scholarly community in Africa have worked closely to address challenges faced across the continent by offering foundational programming and data science workshops to scholars with little or no computational experience. The first Carpentries workshops in Africa were run in 2013, and we have gone on to run 165 workshops (in Botswana, Ethiopia, Gabon, Ghana, Kenya, Mauritius, Namibia, Sudan, Somalia, South African and Zimbabwe as of August 2022), train and certify 125 Instructors and 16 Trainers, and hold two regional community conferences (CarpentryConnect Johannesburg 2018 and CarpentryConnect South Africa 2021) on the continent in the years since.

In November 2019, through support from the South African Center for Digital Language Resources (SADiLaR), Dr Angelique Trusler was onboarded as Regional Consultant for Southern Africa at The Carpentries. In this role, Dr Trusler advanced SADiLaR's community and Carpentries workshops in Southern Africa;  developed a plan to create and support a strong team of volunteer instructors, and developed strategies and workflows to adapt The Carpentries'

[13] Jordan et al, 2018
[14] Barnes, Jordan & Michonneau, 2021

membership and workshop programs to function optimally across the region. Reflecting the rapid growth of the community in Africa, Dr. Trusler took on the role of African Capacity Development Manager for The Carpentries in October 2021. In this role she is responsible for leading The Carpentries efforts to support and grow our volunteer instructor community and member organizations, and to increase our collective impact on researchers across the African continent.

**The Center for Scientific Collaboration and Community Engagement** (CSCCE) champions the importance of human infrastructure for effective collaboration in STEM. We provide training and support for the people who make scientific collaborations succeed at scale (community managers) - and we also research the impact of these emerging roles. CSCCE staff members have expertise in creating and publishing frameworks and resources for community engagement, providing consultancy on a range of multi-stakeholder, open science, learning, and technology-focused projects, as well as developing and delivering training for STEM community builders. We are systems thinkers with expertise in facilitation, program design and project management and our formal training spans the life sciences, social sciences and organizational theory. Our published materials have been downloaded more 31,000 times and we've worked on topics that include metadata standards, expanding the reach of communities internationally, creating team cohesion, and deploying and supporting the adoption of new technologies. This experience will be valuable in multiple ways on this grant - primarily at the strategic level to aid the internal coordination of the multiple team members and to guide the external-facing community-building approach.

**Invest in Open Infrastructure** is a nonprofit initiative dedicated to improving funding and resourcing for open technologies and systems supporting research and scholarship. We advance collective understanding and strategy for the adoption of and investment in open infrastructure through the three core pillars of our work: (1) conducting research and analysis, (2) providing strategic decision-making support to key stakeholders, and (3) engaging with the community to run pilots and trial ways to put our research into practice. Our leading activities to date include: launching the Catalog of Open Infrastructure Services to model a means of standardizing information about core open infrastructure services, and conducting research into the costs of open infrastructure, ways to assess financial health of nonprofit service providers, and good community governance practices for open infrastructure *(to be published in August 2022)*.

## Historical collaboration with biomedical communities

The team on this grant has several historical collaborations with the biomedical community. We will leverage these pre-existing collaborations and successes to build upon this impact for these and other communities in biomedicine. Below we describe several pre-existing efforts between major organizations on this grant.

**The Carpentries** community of Instructors has run 122 Data Carpentry Genomics workshops, teaching skills required to execute a variant calling pipeline on publicly accessible data, to a total of ~2000 learners. Their list of member organizations includes many research institutes that conduct biomedical research, as well as organizations supporting research such as the Bioconductor project and the Software Sustainability Institute. The Carpentries has run 165

workshops in Africa and 53 in Latin America to date (all curricula). They have trained and certified 125 Instructors and 16 Instructor Trainers in Africa and 104 Instructors and 4 Trainers in Latin America.

**MetaDocencia** has contributed to the capacity building of over 1,000 Spanish-speaking participants in 30 countries by teaching 81 workshops since March 2020 and presenting at several online conferences on research, software or education. Part of its impact in the biomedical field includes:

- Teaching for the community of Women in Bioinformatics and Data Science Latin America (WBDS-LA), a community that organizes a number of activities including a yearly conference that unites 500 to 1,000 Latin American Bioinformatics researchers each year.
- MetaDocencia has also built Open Science capacities for the ARPHAI project, one of the grantees around AI and COVID-19 response in the Global South, and this Community of Latin American Bioinformatics Software Developers.
- **CCAD**, a partner of MetaDocencia and collaborator on this proposal, has provided infrastructure used to contribute 82 scientific articles in the biomedical field.[15]

**2i2c** is currently running cloud infrastructure for the NeuroHackademy summer school, which teaches neuroimaging methodology to cohorts of world-wide learners via cloud infrastructure. Co-founding 2i2c team members (PI Holdgraf and founding member Fernando Perez) have spent several years as neuroinformatics researchers and wish to grow their connections to this and related communities.

## Connections and intended collaborations with biomedical communities

Finally, through our previous partnerships and new connections made while designing this project, there are several biomedical communities that we are already connected with, and that we believe will benefit from infrastructure services like the ones described in this proposal.

For example, below we list examples of connections to key communities that already exist for each organization on our “Community Guidance” team (MetaDocencia, The Carpentries, and OLS):

- **MetaDocencia:** Women in Bioinformatics and Data Science Latin America, a cohort of grantees around AI and COVID-19 response in the Global South, this Community of Latin American Bioinformatics Software Developers, and a potential collaboration with Bioconductor to build capacity in Latin America. In addition, CCAD has published nearly 82 articles related to biomedicine using the infrastructure served by collaborators on this grant.
- **The Carpentries:** The Carpentries community includes a large number of biomedical scientists, as demonstrated by the Data Carpentry: Genomics curriculum and the large number of biology-related lessons developed by community members in The Carpentries

[15] Thanks to Mario Alfredo for compiling this list.

Incubator. This connection to biomedical research includes community members who also belong to H3ABioNet, with and through whom we aim to further build connections with this proposal. H3ABioNet wishes to deliver and receive more bioinformatics training, and is interested in utilizing open cloud infrastructure for this. In addition, The Carpentries has been strengthening connections with the Bioinformatics Hub of Kenya (BHKi), including training several Instructors. Note that the team behind that BHKi project was also part of an early OLS cohort.

- **Open Life Science**: As the name may suggest, OLS's team of four directors come from biomedical and bioinformatics backgrounds, and have strong ongoing or historic ties with EMBL, ELIXIR, the Department of Genetics at the University of Cambridge, the Open Bioinformatics Foundation, and the Galaxy bioinformatics workflow platform. When providing leadership training, OLS would draw from their 400-strong community of mentors, experts, and training facilitators. Community members span many prominent biomedical institutes, including EMBL-EBI, The Francis Crick Institute, Biocommons Australia, and research universities worldwide. The community represents networking and expertise in topics from genomics and proteomics, microscopy and imaging, bioinformatics, computational biology workflows, virology, infectious diseases, and drug discovery.

In addition, we have made several connections to groups in Latin America and Africa who are specifically interested in the infrastructure described in this proposal. We aim to collaborate with these groups via OSCS infrastructure services described above. Examples of these communities include:

- Pan-African Bioinformatics Network for H3Africa (H3ABioNet): H3ABioNet was established to develop bioinformatics capacity in Africa, providing training and access to software and data to enable researchers to perform bioinformatic analyses. OSCS resources and training would empower researchers at H3ABioNet nodes to participate in data-intensive research and training in data-intensive methods.
- The Bioinformatics Hub of Kenya Initiative (BHKi): An early beneficiary of OLS training, BHKi takes a community-focused approach to build capacity and create opportunities in bioinformatics in Kenya, through peer training and mentorship. The BHKi community faces obstacles to teach and learn bioinformatics skills due to a lack of access to compute resources.
- CABANA: CABANA is a capacity strengthening project for bioinformatics in Latin America. It aims to accelerate the implementation of data-driven biology in the region by creating a sustainable capacity-building programme focusing on three challenge areas – communicable disease, sustainable food production and protection of biodiversity. The network of researchers that has formed around CABANA includes individuals who would benefit from access to the resources provided by this project, for use in their own research and/or for delivery of CABANA bioinformatics training workshops.
- High-Performance Computing Center in Argentina at the National University of Córdoba (CCAD): provides infrastructure for a variety of research groups in Latin America, and has provided infrastructure for more than 82 scientific articles in the biomedical field.

- One of 2i2c's founding members, Fernando Pérez, has connections with the University of Antioquia (Colombia) with researchers Natalia Gaviria and Juan Rafael Orozco who have expressed interest in this team's cloud computing service. This group performs analysis of biomedical signals in the context of Alzheimer's, Parkinson's and Huntington's diseases[16].

# Risks and mitigation strategies

Below we describe several potential risks that we've identified as part of this proposal, and a few suggested ways to mitigate these risks.

## Ensuring community uptake of service infrastructure

This proposal relies on external communities making the decision to utilize and participate in the cloud services that we describe here. To avoid the possibility that there is heavy *under*-utilization of the infrastructure, we have taken these considerations and plans:

- MetaDocencia (MD) and The Carpentries (TC) both have extensive networks within the Latin American and African communities[17]. They have identified this need and plan to leverage these networks to find interested communities. In addition, we have already identified several interested organizations in this service (see above).
- The "train the trainers" and leadership capacity building aspects of this proposal via MD, TC, and OLS are aimed to grow community buy-in with the goal of leveraging these connections to grow our network.
- 2i2c has already demonstrated a clear interest in these kinds of hosted infrastructure environments via the other communities it works with (it currently serves more than 35 communities in research and education).

## Capacity limits on our team and infrastructure

Another possibility is that there is *over*-utilization and interest in this service, which puts extra burden on our team's ability to deliver the service together. It may also burn through the budget allocated for cloud credits too quickly. Here are major considerations we have taken around this possibility:

---

[16] A longer description Gaviria and Orozco group's work: "Our team works on the development and application of different pattern recognition techniques to model different patterns typically observed by clinicians in several diseases including Parkinson's, Alzheimer's, and Huntington's. Among the signals we have considered (or recently started to consider) to create the models are speech, language, gait, facial movements, and other limb movements. Our main aim is to develop open-source technology for the automatic and non-intrusive evaluation and monitoring of patients suffering from different conditions (like the ones mentioned above). From the technical point of view, currently, our main focus for the near future is the development of methods that enable the synchronous analysis (modeling) of different information sources, e.g., speech production + face movement. This leads us to the evaluation of methods that combine recurrent units with attention layers, gated multimodal units or similar, and others." Current publications from this team: https://sites.google.com/view/rafaelorozco/publications."

[17] For example, here's an overview of MetaDocencia's activities with Latin American communities in the last year.

- First, if we reach our capacity limits quickly (for people or cloud costs), this will be a clear indication of value for these communities, and we hope to be able to leverage this to make a case for more funding via third-party donors, philanthropies, cloud providers, etc.
- We have chosen a use-case to focus on (workshop-based education) that is relatively easier to scale. This tends to use fewer resources and is repeatable across workshops, which will reduce labor and cost associated with each.
- We have historically partnered with organizations that occasionally provide cloud credits in the past (e.g., Google, AWS, NASA, etc). We aim to leverage these connections to procure cloud credits for these communities in addition to the budget provided here, to give ourselves some breathing room.

## Long-term viability of this model

A stated goal of this proposal is to build a sustainable and generalizable *model for a participatory service* that could be brought to other communities and scaled with the communities served in this proposal. What if our research and strategic work makes us realize that this model is fundamentally unsustainable? For example, perhaps the mismatch between salaries paid for personnel in this grant and ability to pay to recover costs via these communities is unresolveable. Here are a few considerations we have taken:

- Understanding the major barriers to sustainability will still be of large value in understanding how to serve these communities in the future. All of the materials and learnings created here will be open and designed for re-use by others.
- The two years of activities under this grant will still serve a broad audience in historically under-served parts of the world.
- We have included a focus on *capacity building* in this grant in order to build core competencies and skills that these communities can bring with them, with or without the team behind this service.

## Collaboration between several different organizations

Although all organizations involved share the vision and mission of this project, we do not have significant experience working together in a project of this magnitude and complexity. How can we ensure that we have appropriate channels of communication, planning, and coordination to make this team “greater than the sum of its parts”? Here are a few considerations we have taken:

- Meet with one another to align on an action plan. Our first step will be to spend time aligning with one another about a shared vision, language, and plan for next steps on the grant. We must also determine our shared understanding and goals for the hand-off points for content creators and workshops, the plan for reaching out to communities and cultivating their growth, and beginning the CCAD secondment with the Cloud Infrastructure team. We will have a number of team and stakeholder meetings and conversations, planned and facilitated by CSCCE - and in partnership with the project coordinator at OLS once they are hired (see next point). This will set up the project for

success from the outset by building the shared understanding and processes necessary to deliver its goals.
- We have dedicated time for a *Project Manager and Coordinator* to provide high-level oversight for this project and oversee efforts to ensure communication and information flow between team members. This person will jointly serve as our community coordinator via a partnership with OLS, which will allow them to have extra context about the service itself and where we are experiencing tensions related to inefficient coordination or planning. We will also have strategic support and guidance from CSCCE and IOI, which have experience guiding large multi-stakeholder collaborations like this one.
- In this proposal, we have established a division of roles between different partners. Leadership and responsibility for each element of the project will be assigned among the partner organizations at the outset of the project.
- Although the organizations have not collaborated on a single project before, ties exist between the organizations and their respective staff and communities, and many members of the project team have a history of working together.
- This project involves a strategic alignment between the long-term plans and goals of each organization, in a way that is complementary and could potentially continue moving beyond this grant[18]. Each organization has an interest in making this collaboration as successful as possible.

## Diversity, equity, and inclusion among cloud service beneficiaries

Community recruitment may include in this program communities that are already privileged within Latin America and Africa. From experience in the territories, we know that local power dynamics often prevent some communities accessing resources like the ones this project will offer.
Here are a few considerations we have taken:

- The Carpentries, MetaDocencia, and OLS are organizations that recognize this challenge, and are genuinely committed to diversity, equity, and inclusion and will work to design fair onboarding to this program.
- Centering marginalized voices: in scenarios where personal or organizational experiences differ, all teams agree to center the experiences of marginalized experiences over privileged experiences, and to offer flexibility for differing circumstances when values or needs may conflict.

[18] For example, this work aligns with Goal 2 of The Carpentries Strategic Plan for 2020-25, to "Intentionally incorporate equity, inclusion, and accessibility to support a diverse community. It also aligns with 2i2c's stated need to find a better sustainability model around services like these for under-resourced communities.

# Appendices

## Appendix: Shared Responsibility Diagrams

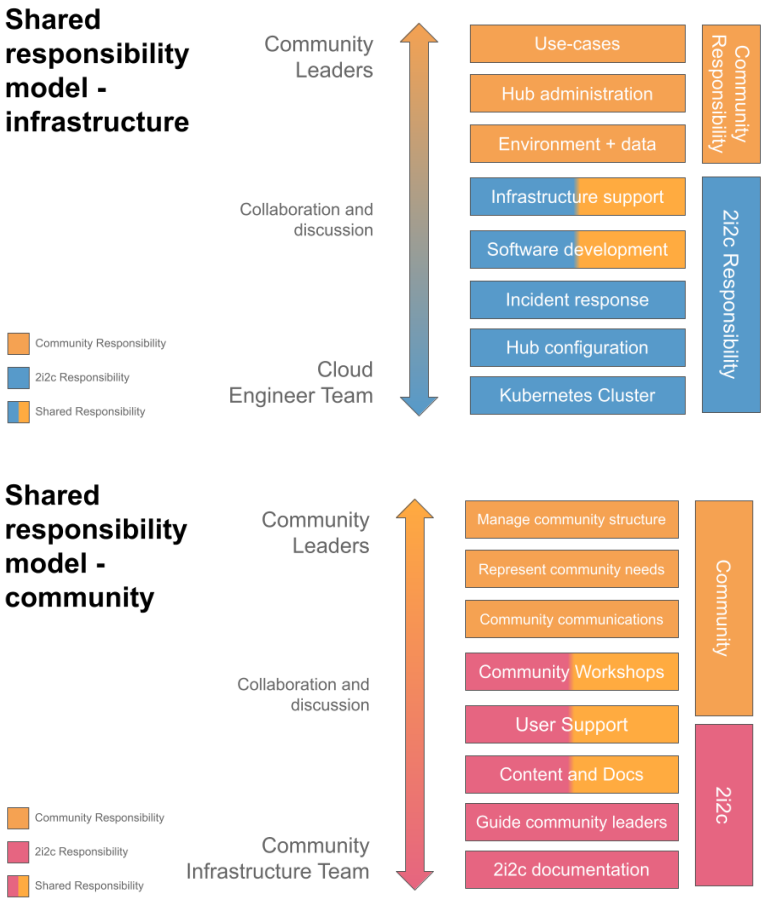


***Figure A1.*** *Our shared responsibility model for infrastructure and community engagement. Running collaborative cloud services requires understanding the organization that is primarily responsible for major activities needed to run the service. We draw inspiration from major cloud providers, which use Shared Responsibility Models to make this breakdown explicit. This usually does not have a "hard boundary" between any two responsibilities, and is dynamic and community-dependent. These figures show our starting points, but we hope to refine and evolve this as we learn more.*