Open Innovation for Earth Observation Programmes

2-4 November 2022 | ESA-ESRIN | Frascati (Rm), Italy

Perspective on creating Open Source user workflows
The Pangeo Community

Anne Fouilloux, Simula Research Laboratory (Norway), annef@simula.no, @AnneFouilloux, https://github.com/annefou

Outline

- What is Pangeo?
- How does Pangeo foster Open Innovation?
- Successes, Challenges and Lessons learned
- Recommendations to Space Agencies

What is Pangeo?

A global community initiative for Big Geoscience Data that promotes open, reproducible, and scalable science

- Open Community
- Open Platform

Code of conduct

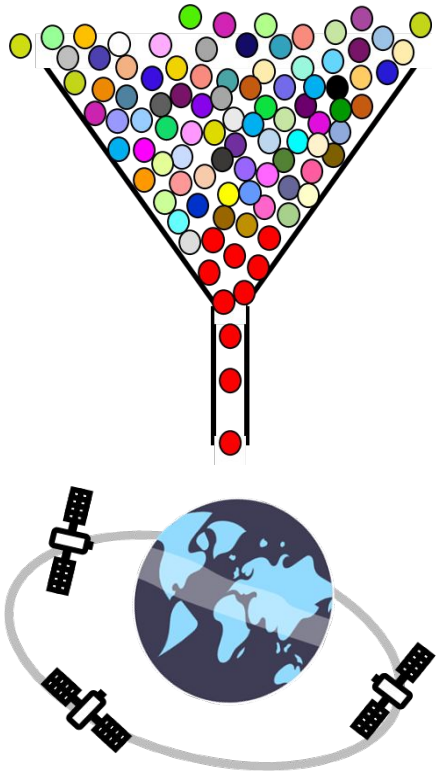
- https://github.com/pangeo-data/governance/blob/master/conduct/code_of_conduct.md

Governance

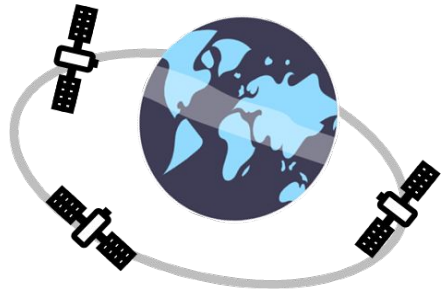
- https://github.com/pangeo-data/governance/blob/master/governance.md

Open Community

Pangeo vision: lower barriers to entry to stimulate innovation



Facilitate contributions to increase **diversity**



DEI (Diversity, Equity and Inclusion) is essential

A community of developers, scientists and users

America
Time zone +11-> -2

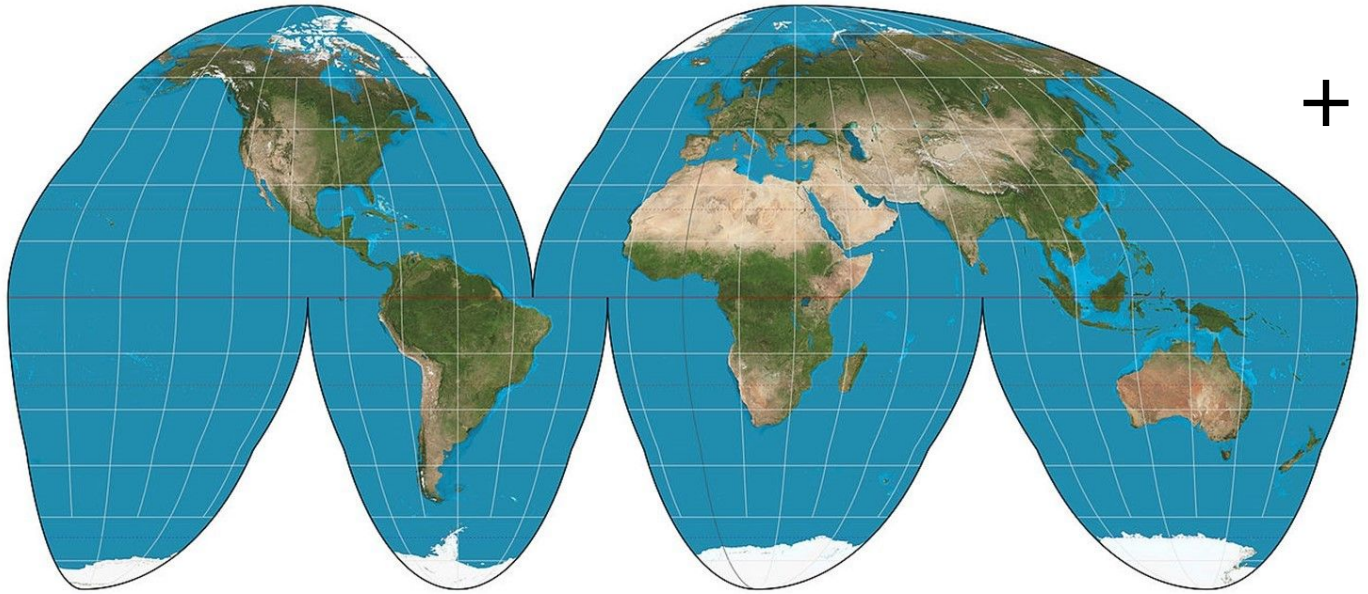
Every Wednesday, alternating between 12pm ET and 4pm ET

Europe, Africa, West Asia
Time zone -1-> +5

Every Tuesday at 9.30 a.m. CET/CEST here

Australia, East Asia
Time zone -1-> +5

3rd Friday of every month at 1pm Australian Eastern Time



+ Working Group Meetings

- *Machine Learning Working Group*
- *Cloud Operations Working Group*
- *Project Pythia (formerly the Education Working Group)*
- *Pangeo Forge (Cloud Data Platform)*
- *Open Science Meeting - discussions to coordinate open science activities*

More info & links here - https://pangeo.io/meeting-notes.html

Pangeo Showcase Webinar Series

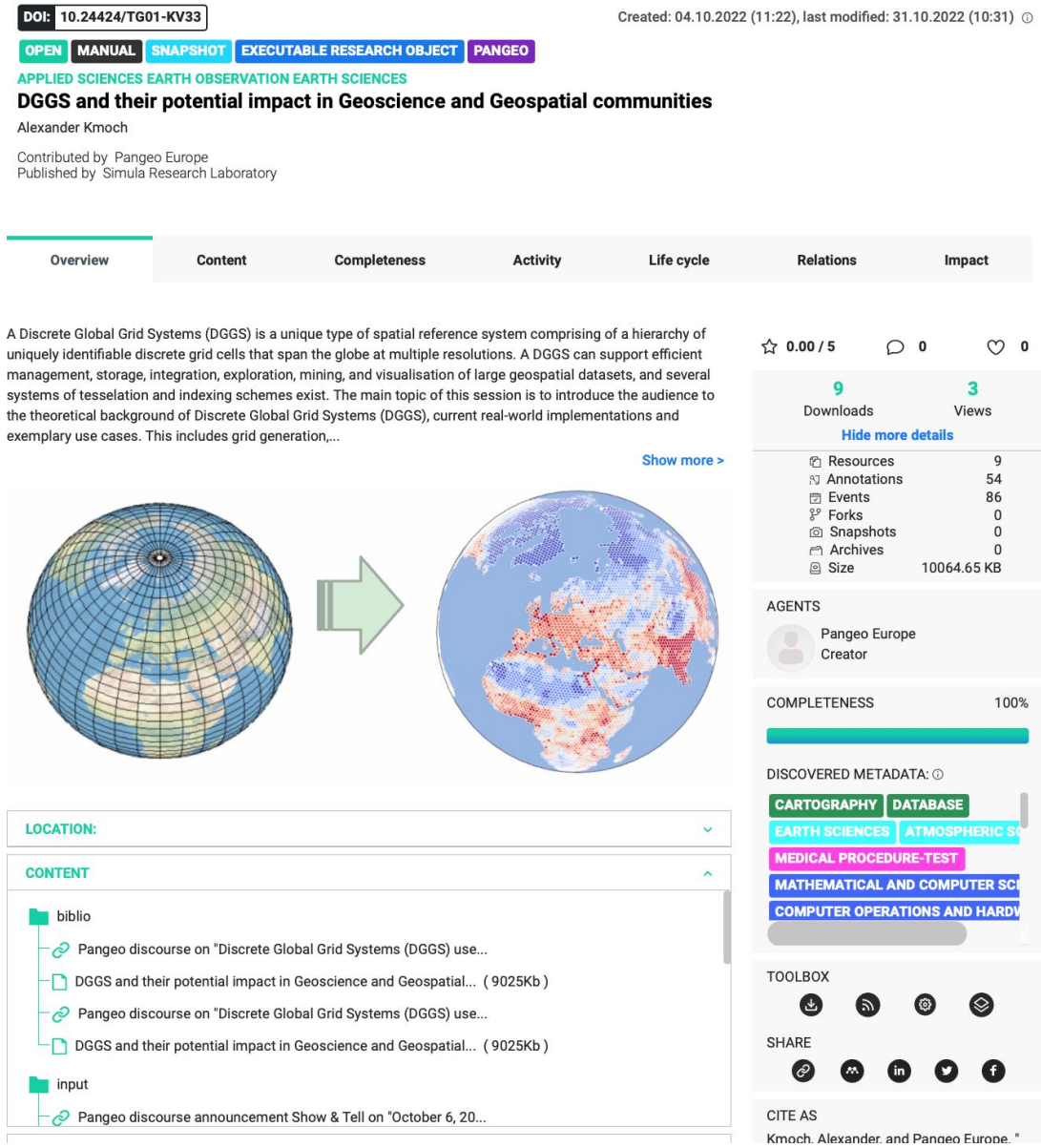
- To learn from each other. It it part of “America” community calls. 15 minute talks are recorded, given a DOI and made available on the Pangeo YouTube Channel;

FALL 2022 SHOWCASE

Date	Speaker	Title
2022-09-21 12PM EDT	Peter Marsh, University of Cape Town	Accessing NetCDF and GRIB file collections as cloud-native virtual datasets using Kerchunk DOI 10.5281/zenodo.7140825
2022-10-05 12PM EDT	Leah Wasser, pyOpenSci	PyOpenSci DOI 10.5281/zenodo.7158586
2022-10-12 4PM EDT	Rich Signell, USGS	My ERA5 Journey: From API-to-ARCO DOI 10.5281/zenodo.7226344
2022-10-19 12PM EDT	Matthias Mohr, openEO	openEO: What it is and how it relates to Pangeo DOI 10.5281/zenodo.7229398
2022-10-26 4PM EDT	Hauke Schulz, CICOES/University of Washington	Xbitinfo: Compress datasets based on their information content

Show & Tell

- At the request of anyone who is interested to **go in depth and show what they are working on**. It is expected to be hands-on and is also recorded.
- FAIR Digital Object: a Research Object (https://reliance.rohub.org/) is created and aggregate all the material used/showed (slides, recorded video, jupyter notebooks or other codes, input data, etc.) and given a DOI.
- Text mining service to extract metadata and facilitate interdisciplinary re-usage



Open Innovation for Earth Observation Programmes

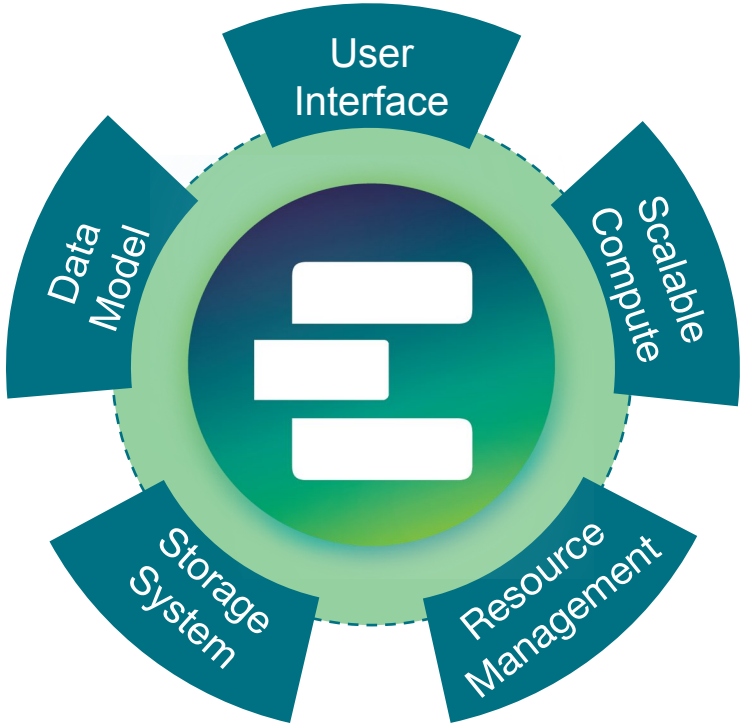
2-4 November 2022 | ESA-ESRIN | Frascati (Rm), Italy

A Culture of Collaboration

Based on GitHub and papers affiliations

Open Platform

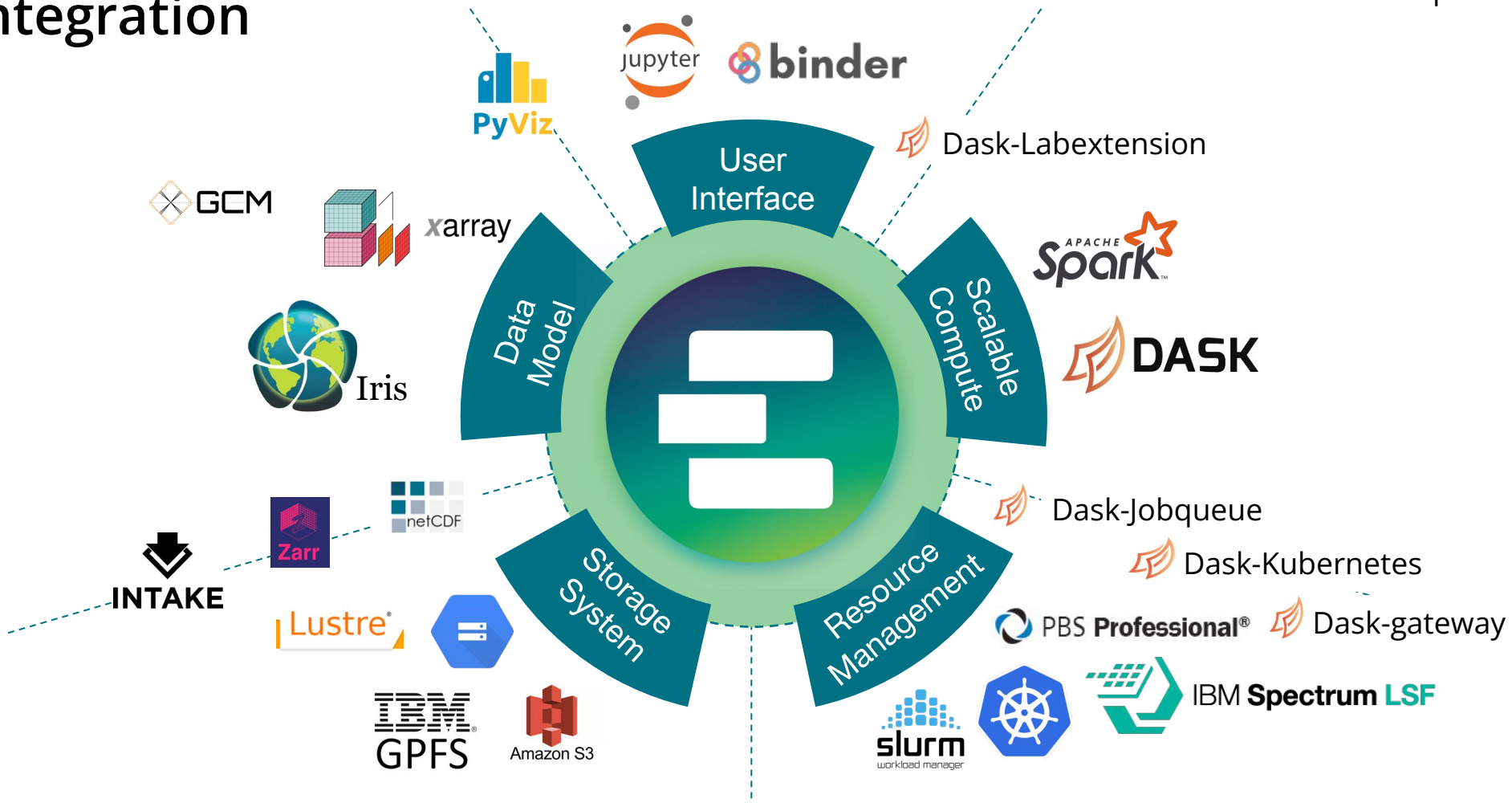
Pangeo **platform** *is scalable*



from laptop to cloud or HPC

Pangeo is about **integration**

Conda integrates most platforms



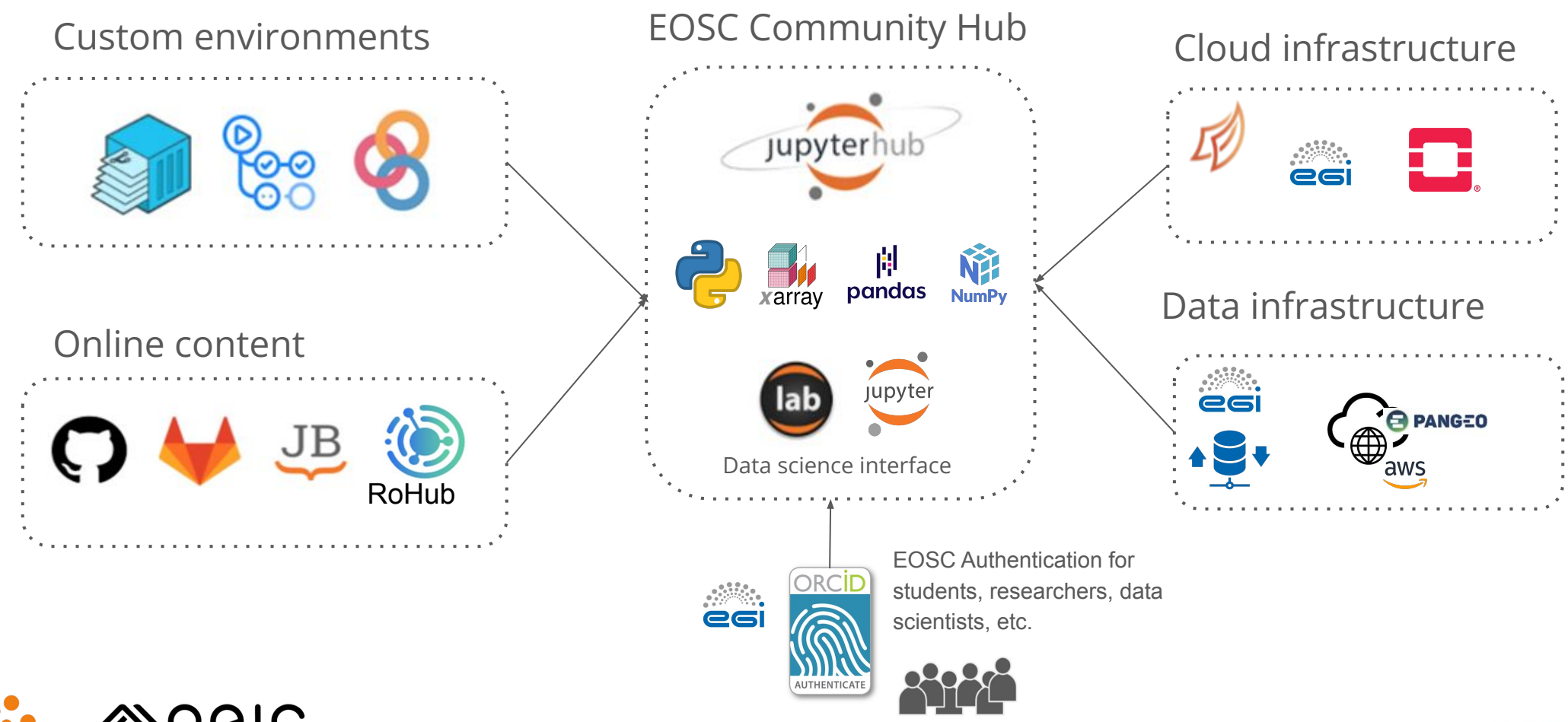
Pangeo makes it possible to explore geoscience data using HPC or cloud in an interactive manner

Pangeo deployments

- HPC
  - CNES, IFREMER, PRACE, Fugaku, ..
- Cloud
  - Public Clouds (EGI-ACE, EOSC)
  - Commercial clouds (AWS, Microsoft, Google)
- Laptop
  - Anywhere for anyone!

Docker images (pangeo-notebook, ml-notebook, etc.) developed and maintained by the Community to ease usage and deployments.

Pangeo at scale with deployment on the European Open Science Cloud (EOSC)



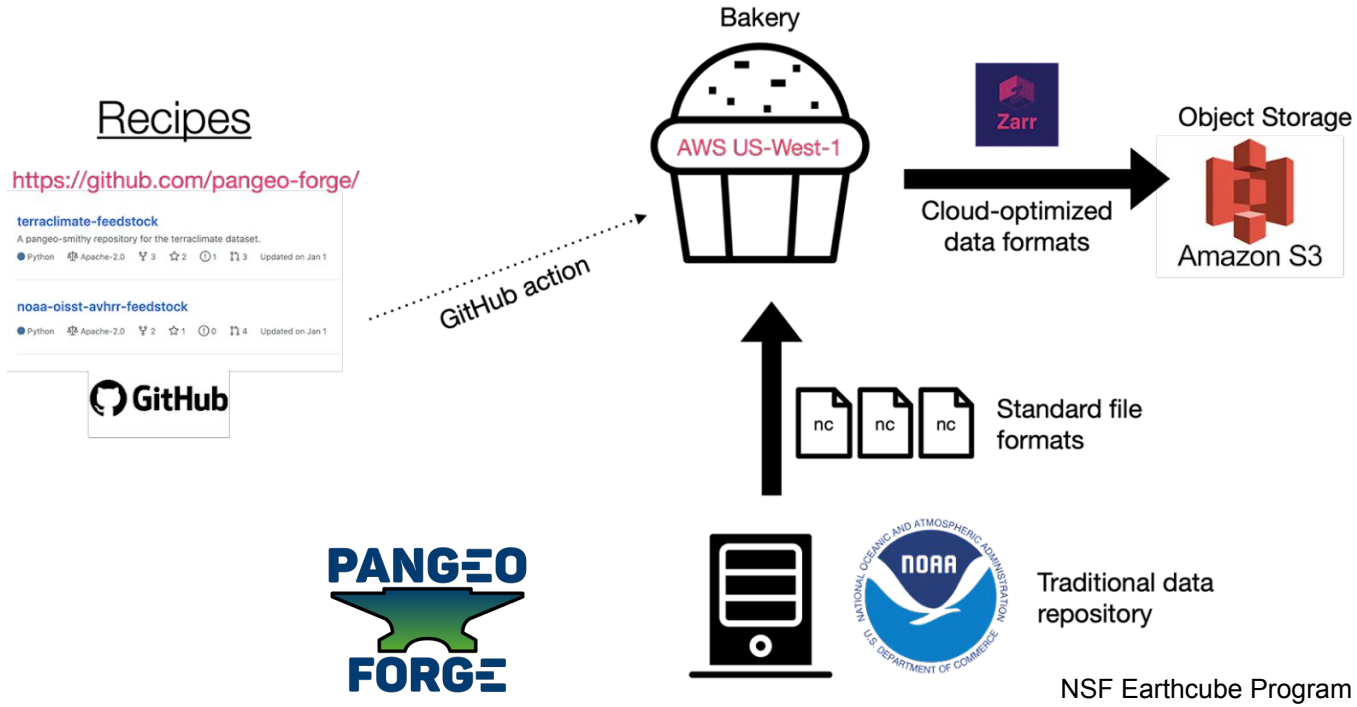
Easy and efficient access to Geoscience data

Pangeo-forge for efficient data management within the Pangeo Community

→ Make it easy to extract data from any traditional repository and deposit this data in cloud object storage in an analysis-ready, cloud optimized (ARCO) format;

Two main components:

- Open source Python package for describing and running data pipelines;
- Cloud platform for automatically executing recipes stored in Github repos.



https://pangeo-forge.org/

STAC and Pangeo

- Pangeo-forge supports the creation of analysis-ready cloud optimized (ARCO) data in cloud object storage from "classical" data repositories;
- STAC is used to create catalog and goes beyond the Pangeo ecosystem;
- Work is ongoing to figure out the best way to expose Pangeo-Forge-generated data assets via STAC catalogs.

PANGEO

STAC
SpatioTemporal
Asset Catalog

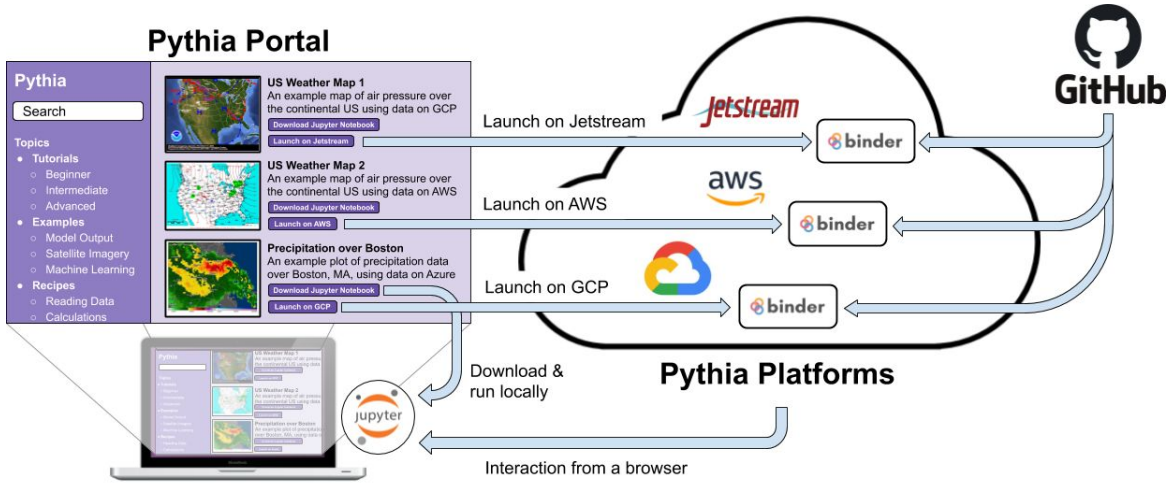
Team-up with other initiatives to onboard new members and increase diversity

Team-up with Project Pythia

A Community Learning Resource for Geoscientists

https://projectpythia.org/

This work supported through the National Science Foundation Award #2026899.



Aspiration goal: Be the goto resource for learning the *Scientific Python Ecosystem*

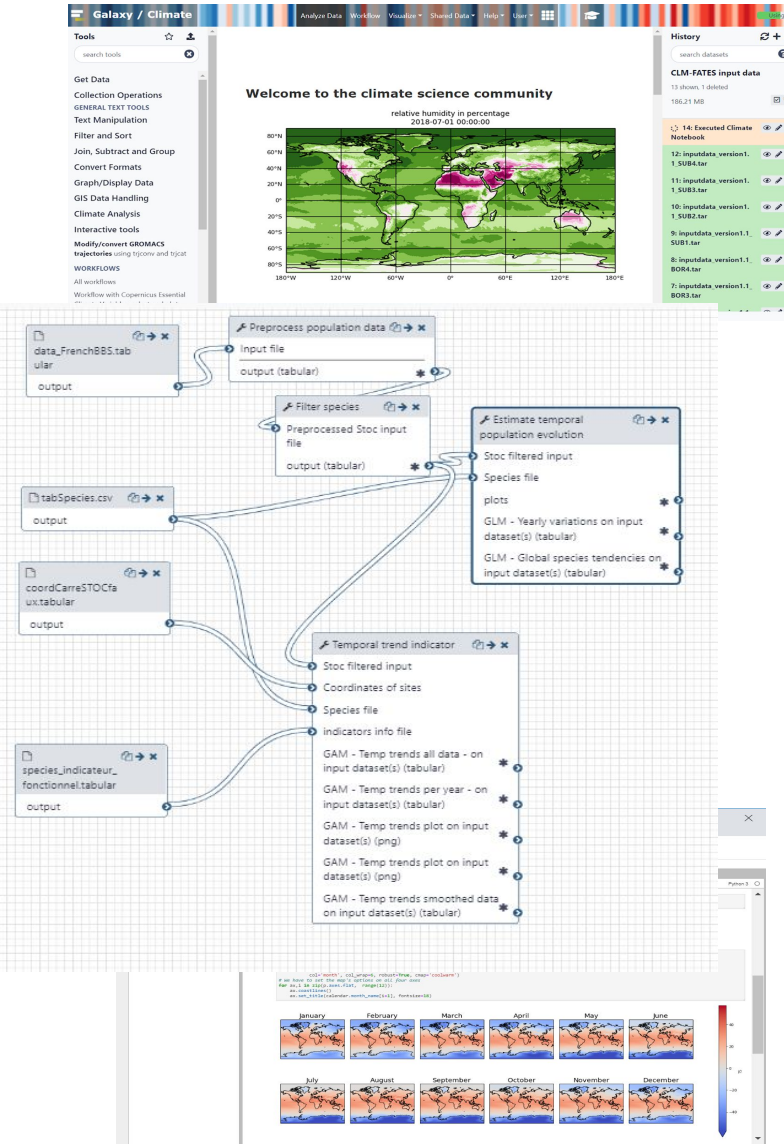
- ★ Geoscience focused
- ★ From beginner to the power user
- ★ Tutorials, videos, examples, on-line courses, and sample data
- ★ Community owned!

Source: Clyne et al (2022). ARM/ASR Open Science Virtual Workshop May 10, 2022

Team-up with Galaxy Europe

Galaxy is an open-source community and platform for FAIR data analysis. It offers:

- **Graphical User Interface (GUI) for users with no programming skills**
- Workflow editor to create and run fully reproducible data analysis
- Compute & Storage to everyone (free registration)
- Self Paced Learning material with the Galaxy Training Network
- Training Infrastructure as a Service (TiaaS) free and ready to use with private queues where only training's jobs run



Galaxy PROJECT

Co-design: from Jupyter Notebook to standalone Galaxy Tools

Create a Jupyter Python/R/Julia Notebook

Galaxy Europe

One Jupyter Notebook is turned into one script

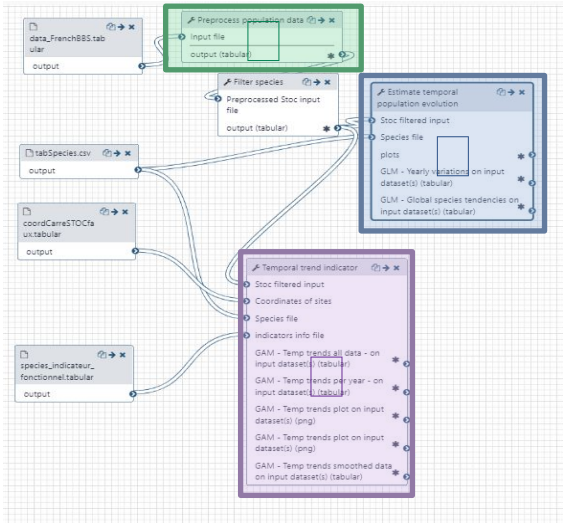
```
                ,direction="wide")
  tab[is.na(tab)] <- 0
                                  # filename <- "touverUnNom"
                                  #  chemin <- paste(rep,filename,sep="/")
                                  #  write.table(tab, chemin)
colnames(tab) <- sub("nombre.","",colnames(tab))

  return(tab)
}

## sous jeux de donnees si choix d espece d annee ou d un pourcentage de carres
makeSousTab <- function(tab,vecSp=NULL,echantillon=1,
                        methodeEchantillon="carre",vecannees=NULL) {
  cat("  -- Fabrication du sous jeu de donnees --\n")
  flush.console()
  ## reduction de la table à certaine espèces
  if(!is.null(vecSp)) {
      cat("      selection",length(vecSp),"espece(s):\n -> ")
      cat(vecSp)
      cat("\n")
      tab <- data.frame(carre = tab$carre,annee = tab$annee,tab[,vecSp])
      colnames(tab) <- c("carre","annee",vecSp)
  }
  ## reduction de la table pour certaines annees
  if(!is.null(vecannees)) {
      tab <- subset(tab,annee>=vecannees[1] & annee <= vecannees[2])
  }
  if(echantillon != 1) {
      if(echantillon < 1 & echantillon >0) {
          nbinit <- nrow(tab)
          if(methodeEchantillon == "global") {
              nb <- round(nrow(tab)*echantillon)
              cat("     echantillonage",echantillon*100,
                  "% des donnees par la methode",methodeEchantillon,"\n")
              cat(" -> conservation de",nb,"lignes sur",nbinit,"\n")
              flush.console()
              tab <- tab[sample(1:nrow(tab))[1:nb],]
          } else {
              if (methodeEchantillon =="carre") {
                  cat("     echantillonage",echantillon*100,
                      "% des carrees par la methode",methodeEchantillon,"\n")
                  nbcarreinit <- length(unique(tab$carre))
                  chat=sample(unique(tab$carre),
                      length(unique(tab$carre))*echantillon,replace=F)
                  cat(" -> conservation de",length(chat),"carrees sur",
                      nbcarreinit)
                  tab=subset(tab, subset = carre %in% chat)
                  cat(" (",nrow(tab)," lignes sur ",nbinit,")\n",sep="")
              } else {

                  stop("Methode d echantillonnage non reconnue")
```

Several atomic scripts from which generic and fully annotated Galaxy Tools are created

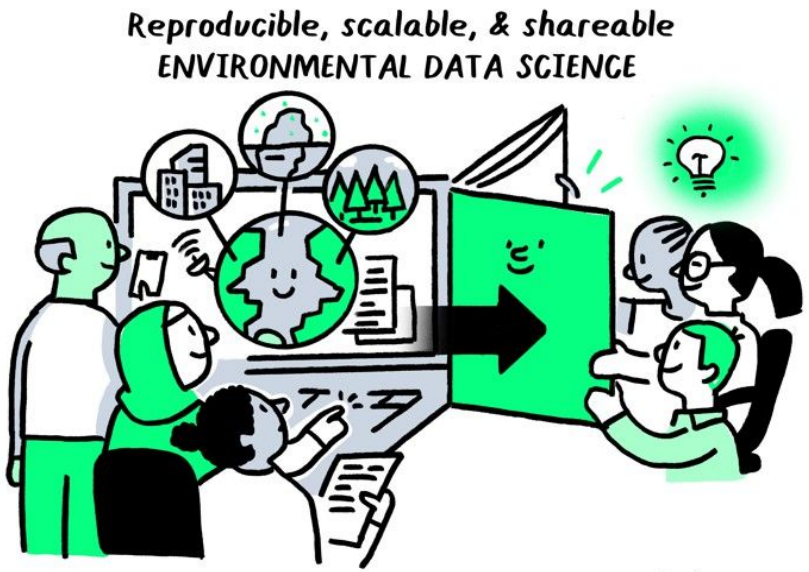


Slide adapted from Coline Royaux, Natural History Museum, France.

Team-up with the Environmental Data Science Book

https://the-environmental-ds-book.netlify.app/

*Living, open and community-driven online resource to **showcase and support the publication** of data, research and open-source tools.*



The Alan Turing Institute

Scriberia

Contribution

Collaborative Reviewing

Publication

jupyter {book}

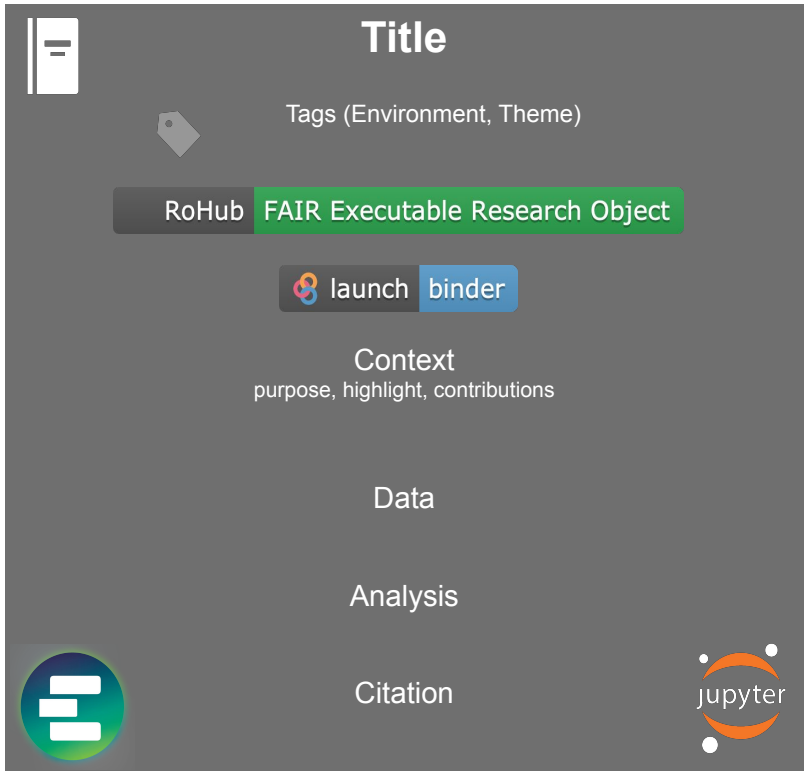
binder + jupyterhub

zenodo doi

EXPLORE

Reliance

RoHub



Initiative led by Alejandro Coca-Castro

Successes

- **No vendor lock-in;**
- Easy to start and **deployment on laptops, cloud and HPC;**
- “Reference” deployments on different cloud infrastructures;
- **Team up with other initiatives.** It can help to increase DEI (Diversity, Equity and Inclusion):
  - Educational material and deliver trainings (Pythia, Galaxy);
  - Training infrastructure as a Service (Pythia, Galaxy, EOSC);
  - Use Pangeo from GUI (no programming skills required) on Galaxy Europe.
- **Contribute to easy creation of data in analysis-ready, cloud optimized (ARCO) format (pangeo-forge);**
- Promote the work done by the Pangeo Community and other Geosciences initiatives (Pangeo Show & Tell/Showcase);
- Pangeo heavily used in industry;
- **Spin-off** (often from Pangeo community members) and many startups & companies using Pangeo software stack and contributing to Pangeo ecosystem.

Challenges

- **Enthusiastic individuals and volunteered driven community:** Pangeo is growing in many different ways and it is often difficult to inform everyone about all the initiatives;
- Initiatives are usually driven by a few dedicated individuals with short-term funding (or no funding);
- Several **spin-off** but not easy to track them down and/or help them;
- Difficult to decide when a sub-project needs to be "professionalized" e.g. being able to "collect" funding, secure long-term funding (NumFOCUS but US based only);
- Difficult to onboard (train) wide range of contributors;
- **Roadmap is community driven** and can be a bit obscure to newcomers;

Lessons learned

- Important to create an Identity early on ;
- Need to team-up with other initiatives for instance to get infrastructure, training network, interdisciplinary research, and different types of contributors (**diversity of contributions**), etc.;
- Important to be able to bootstrap new sub-projects (new packages);
- Need to fund a **Community Manager.**

Recommendations to Space Agencies

Space Agencies could play a larger role, and in particular:

- Contribute actively to Open Source and Open Science projects (prevent re-inventing the wheel);
- Be more involved in Open Science initiatives, and even lead some of them (cf. #NASATOPS);
- Provide infrastructure (compute & Storage)
- Sponsor Open Source projects (similar to NumFOCUS) with EO industry compliance;
- Make calls to attract “small-players” and foster innovation;
- Fund training/hackathons events to engage with a wider range of actors, including training of community manager (CSCCE);
- Promote and fund initiative such as OSGeo;
- Support and invest in tech start-ups (incubators).

What about a “Matchmaker program” to connect with existing/upcoming entrepreneurs, deep-tech consultants, researchers, community managers, software developers, governments, citizen, funders, etc.?

Thank you for your attention!