NFDI4Earth

Deliverable D1.2.6

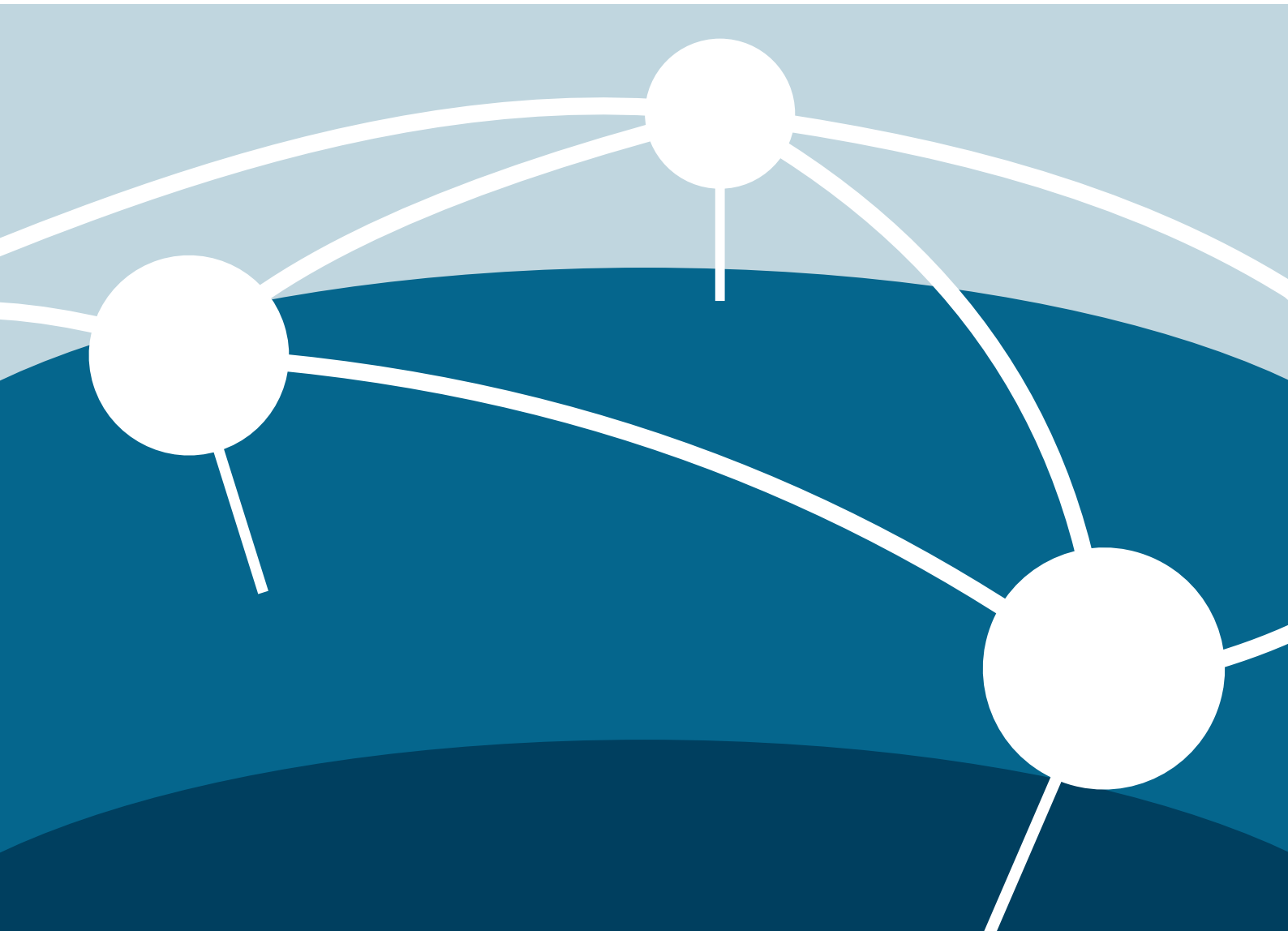
Report and Evaluation of Incubator Lab Projects

Steffen Busch (Steffen.Busch@ikg.uni-hannover.de)

2023-06

DOI: 10.5281/zenodo.13837883

nfdi4earth.de



Citation

Steffen Busch. 2024. *Report and Evaluation of Incubator Lab Projects (NFDI4Earth Deliverable D1.2.6)*. NFDI4Earth Community on Zenodo. https://doi.org/10.5281/zenodo.13837883

Acknowledgement

This work has been funded by the German Research Foundation (DFG) through the project NFDI4Earth (DFG project no. 460036893, https://www.nfdi4earth.de/) within the German National Research Data Infrastructure (NFDI, https://www.nfdi.de/).

Executive summary

This document summarizes the Incubators' results and the evaluating of the resulting deliverables, by validating their GitLab repositories and reports. The projects are evaluated along a set of criteria, as defined in section 3. In addition, interviews with the researchers of the projects are conducted in order to evaluate their own assessment of their achievement and also the whole process. This assessment and evaluation is part of an ongoing monitoring process, which also will be directly exploited for NFDI4Earth *trend scouting*, as well as for identifying potential further research areas for the whole community. In this way, we potentially provide opportunities for follow-up projects and the participation in interdisciplinary collaborations.

Contents

1 Overview

The four Incubator projects of the first call started with an online Kick-off event in August 2022, followed by individual meetings with the coordinator in September. On 19th and 21th October the Incubators presented together with the Pilots their mid-term presentation, followed by individual coordinator meetings in December to structure the final report and clarify the deliverables. On the occasion of the NFDI4Earth plenary in Dresden on 2nd June, two Incubator projects, HDF4Water and scrAIber, were presented.

2 Evaluation Criteria

In the following, the Incubator projects are evaluated along the FAIR principles and also with respect to their potential for follow up projects or inclusion in future NFDI4Earth services. The evaluation relies on a checklist of criteria and an interview with the researchers who conducted the projects, to provide an assessment of usability.

- Findable
  - repository findable via DOI (is your project listed as NFDI4Earth Incubator at Zenodo)
  - completeness of repository, check if all files are present (see README.md template)
- Accessible
  - completeness of Citation.cff (see README.md template)
  - (optional) if a service is provided, check its availability
  - (optional) if other repositories are used, check them against > this FAIR criteria
- Interoperable
  - verification of Jupyter notebook functionality (at best *Binder* > ready)
  - verification of references to meta data of the data and > repositories used
- Reusable
  - verification of the LICENSE file
  - (optional) verification of consistency of license of used > repositories
  - (optional) verification of consistency for license of used > services
- Suitability
  - comparison of proposal promise and deliverables

  - verification of publications

- Utility
  - verification of transferability to other disciplines
  - verification of outlook (follow-up projects or potential > applications)
  - verification of citations (publications, Zenodo DOI, repository > stars, repository forks, repository last activity)
  - assessment of community response (search for follow-up projects)

In the following, all projects are validated against the evaluation criteria from section 3 and a brief (subjective) assessment of their future potential is provided.

2.1 IPFS Pinning Service for Open Climate Research Data

In this project a IPFS Pinning service was implemented which provide a reliable and trusted way for users to ensure that their files remains accessible, because of an distributed redundant file system. This project is a prove of concept of the IPFS pining service in the context of climate data, and to show how the pinning service could also used locally.amssymb This project seems have good chances for follow-on project due to its general interdisciplinary potential. Furthermore the high overall evaluation score (see Tab. 1) ranks it as an success story for an technology transfer. A publication at the EGU General Assembly 2023 conference has already publicized the progress of this project in the community.

Table 1: Evaluation IPFS Project

Criteria	Fulfils	Comments
Findable		
DOI	x	10.5281/zenodo.7646355
Findable	x	By *Incubator* keyword
Completeness	x	
Accessible		
Complete cff	x	Research field is missing
Service	-	Prove of concept, may run service for costumers
Repositories	-	
Interoperable		
Jupyter	x	Binder ready
References	-	
Reusable		
License	x	MIT

Criteria	Fulfils	Comments
Repositories	-	
Data	-	
Suitability		
Deliverables	x	IPFS service is provided
Publication	x	Abstract at EGU2023, https://doi.org/10.5194/
Utility		
Transferability	x	Universal filesystem
Outlook	x	General interdisciplinary potential
Citation	0	
Stars	0	
Forks	0	
Activity	4 month	
Response	medium	Implementation for climate researchers

2.2 ScrAIber

This project implements a functional web service for a specific target group, namely geologists or more specifically mineralogists. This service segments olivine crystals in back scattered electron images. The general project idea of offering AI segmentation as a service to people without programming skills seems to have great potential. However, the current realization is limited due to the special application and data types involved.

Table 2: Evaluation ScrAIber project

Criteria	Fulfils	Comments
Findable		
DOI	x	10.5281/zenodo.7744225
Findable	x	By *Incubator* keyword
Completeness	(x)	Missing project description
Accessible		
Complete cff	x	
Service	x	Freely usable service
Repositories	-	
Interoperable		
Jupyter	x	Deployed by Docker container ready
Meta data	-	
Reusable		
License	x	MIT-0
Repositories	-	

Criteria	Fulfils	Comments
Data	-	
Suitability		
Deliverables	x	
Publication	-	Implementation of existing technology
Utility		
Transferability	-	Fix data format and segmentation
Outlook	-	No project description
Citation	0	
Stars	0	
Forks	0	
Activity	4 weeks	
Response	low	Implementation for mineralogists

2.3 HDF4Water

In this project, HDF5 has been successfully used for efficient use of Open-Street-Map (OSM) water data in AI frameworks and for efficient visualization of the data. This is done by indexing OSM data to fit the entire world in memory. The conversion of OSM vector data to HDF5 provides an interface for easy access, in terms of analysis-ready data, for common GeoAI frameworks. A pub- lication in the Proceedings of the 10th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data already made the progress of this project known in the community.

Table 3: Evaluation HDF4Water project

Criteria	Fulfils	Comments
Findable		
DOI	x	10.5281/zenodo.7562586
Findable	x	By *Incubator* keyword
Completeness	x	
Accessible		
Complete cff	(x)	no references: nfdi4earth community and research disciplines (zenodo subjects)
Service	-	
Repositories	x	
Interoperable		
Jupyter	x	Deployed Docker container
Meta data		
Reusable		
License	x	MIT

Criteria	Fulfils	Comments
Repositories	x	consistent license
Data	-	
Suitability		
Deliverables		
Publication	x	https://doi.org/10.1145/3557917.3567615; BigSpatial 22, second author
Utility		
Transferability	x	universal vector conversion
Outlook	(x)	Within the paper, extension for heterogeneous data.
Citation	0	
Stars	3	
Forks	1	
Activity	6 month	
Response	high	Implementation for marine data science, extension for remote sensing, may used by *nextbillion.ai* and *Grab* company.

2.4 Aquatic Ecosystems

This project makes machine learning approaches in image analysis accessible to domain specialists without programming expertise by using LableStudio, and develops its potential within a bigger project. The work was published in a journal article. Unfortunately this article is still under review, so the results and the code publication are still pending. The detailed description of the dependencies, with detailed installation instructions, suggests a valuable contribution to the community at the end of the publication process.

Table 4: Evaluation Aquatic Ecosystems project

Criteria	Fulfils	Comments
Findable		
DOI	x	10.5281/zenodo.7763865
Findable	-	By title
Completeness	-	no videos, pending for journal publication
Accessible		
Complete cff	(x)	missing identifiers and keywords
Service	?	Pending...
Repositories	?	Pending...
Interoperable		
Jupyter	x	not Binder-ready
Meta data		
Reusable		

Criteria	Fulfils	Comments
License	x	GPL-3.0
Repositories	?	Pending...
Data	-	
Suitability		
Deliverables	?	Pending...
Publication	?	Pending...
Utility		
Transferability	x	universal labeling process
Outlook	x	Relevant groups biologists, ecologists, image analysts
Citation	?	Pending...
Stars	0	
Forks	0	
Activity	3 month	
Response	low	Implementation for oceanographers, especially plankton researchers

3 Staff Interviews

Project staff was asked the following questions, both to identify potential follow-on projects and to provide feedback on incubator project coordination.

3.1 Questions

- Are there already follow-up projects in which the output of your projects is used?
- Do you use your knowledge from the *NFDI4Earth Incubator* projects in other projects, or did you turn to other topics?
- Did researchers from other disciplines show special interest in your work; if so from which research areas did they come?
- Did partners from the industry show special interest in your topic?
- Which was the greatest difficulty in the realization of your work?
- How can we support you even better?

3.2 Answers

The project staff's responses are summarized in one bullet point per question.

3.2.1 IPFS – Marco Kulüke

- Marco Kulüke is now part of the ESIWACE project (https://www.esiwace.eu/). This project includes also an IPFS working package in the 3rd project phase, were he will support an IPFS service implementation.
- At the moment, Marco is focusing on the FAIR digital objects, but he will return to IPFS in the later phase of the project.
- Mean response to conference poster from climate change researchers.
- A representative from Intel was curious about the IPFS user experience.
- The short project time was the biggest problem, especially in the preparation of the proposal. It was difficult to formulate the work steps as precisely as the short project duration required.
- Overall satisfied with the support, especially during the project implementation; satisfied with the smooth communication.

3.2.2 ScrAIber – Artem Leichter

- The service is maintained by Artem Leichter for mineralogists. In addition, an expansion is planned, eventually within the framework of a further incubator.
- The software base will be used for further specific segmentation projects.
- The interest of fossil researchers in scrAIber points to further areas of application in biology.
- So far, no industrial interests.
- The main difficulty is adapting the segmentation to different proprietary file formats for specific sensors,the lack of standards for back-scatter-electron microscopes.
- Publicly effective events to promote awareness of the project.

3.2.3 HDF4Water – Hao Li

- Hao Lie use the developed converter for OSM data to AI frameworks by HDF5 file format for a new project, which detect missing building in OSM data. He is also applying for funding for a project to further develop the incubator's overarching *AtlasHDF* project as an IO component for an even larger *Helana* framework being developed at the Technical University of Munich.
  - **The Helena Big Geospatial Data Framework (CLI)** is the swiss army knife tool for AtlasHDF and the surrounding suite of GeoAI tools. This development version of Helena contains several projects: 1) AtlasHDF in which we design a novel spatial storage concept with better HPC properties; 2) Flow in which we implement a high-

level graphical programming language for spatial data science; 3) Search in which we implement various aspects of spatial search.

- Researchers from the discipline of marine data Science plan to use the HDF4Water within a new graduied school in Bremen. Furthermore, there is interest from remote sensing re- searchers from Mu¨nster to complement the developed converter with image data to analyze satellite images. The audience of geoinformaticians and data scientists at the BigSpatial workshop, where the publication was presented, noticed the project as a solid example that shows the right way in the complex handling of data. Thus, the project also shows potential for an EduPilot.
- The Nextbillion (https://nextbillion.ai/) and Grab (https://www.grab.com/sg/) com- panies are checking whether the HDF4Water meets their requirements.
- The biggest difficulty was the short run time of the project, especially to start directly with the approval and to deliver the results in time. Also it was confusing to combine GitLab and GitHub repositories.
- Early access to the deliverables template to clarify the format of the deliverables to be delivered.

3.2.4 Aquatic Ecosystems – Ankita Vaswani

- Several follow up project are already planed and started with in the Helmholz.AI group for the next 1.5 years. Each of the following project use the knowledge, software and label data generated by this incubator project.
  - Instance Segmentation and classification of plankton
  - Latent space anomaly detection for identifying more relevant data for labeling and training
  - Web GUI for easy access to complete pipeline: labeling, training, segmentation, classi- fication ans anomaly detection
- Ankita Vaswani is involved in the above projects and is currently working on a publication on latent spatial anomaly detection based on the label provided by this incubator project.
- In addition to underwater imagery, this software is also of interest for remote sensing of plankton blooms to predict changes in the ecosystem.
- No interest from industry yet.
- The short run time is a problem, especially when delivering software and data according to FAIR principles. It is difficult to place the results of the incubator in the context of a larger project, as it is difficult to create a story for a sub-project without considering the overarching aspects. For this reason, the project is not finished now.
- An NFDI repository steward to provide ongoing support for compliance with FAIR princi- ples.

4 Final Evaluation

After a brief summary of all incubators and their deliverables, an outlook is given by summarizing the experiences of the first incubator cohort.

4.1 Summary

In terms of deliverables, almost all projects were successfully implemented and completed, with only the Aquatic Ecosystems project being only partially completed due to an outstanding journal publication related to the parent project. All projects are findable via a DOI, accessible by public repositories with meta data and reusable by clear license, but they differs in the interoperability. In particular, IPFS and scrAIber have high interoperability through easy access to services and Jupyter notebook tutorials. In contrast, HDF4Water and the Aquatic-Ecosystem project are more complex as part of larger projects. For high interoperability, *ScrAiber* and *HDF4Water* provide a Docker container for using their software. Additionally, *ScrAiber* provides easy access to its Jupyter notebook via an additional Docker container, also the *HDF4Water* project plans to extend its Docker container with Jupyter notebook access. Final interoperability validation for the Aquatic Ecosystem project is pending due to ongoing publication in a peer-reviewed journal. The *IPFS* and *HDF4Water* projects have published and presented their results at conferences and workshops beyond the scope of NFDI4Earth. Compared to the *ScrAIber* and *Aquatic Ecosystem* projects, which are targeted at *mineralogists* and *plankton researchers*, the *HDF4Water* and *IPFS* projects reach a broader audience, including industry, because of their universal applicability.

4.2 Lessons learned

The project focused on small, tangible projects and could be finished within the given time-frame - thus the goal of the Incubators can be considered as being achieved. However, there is also room for improvement: on the structural side, the requirements of the deliverables might have been too demanding. In particular, e.g. despite the template for the Citation.cff, which collects all necessary meta data, the inconsistency and incompleteness of the file shows that it has not been fully understood. For the next incubator cohort, the template will be improved, or a web formula will be used to capture the metadata. In addition, the incubator repository of results will be available as projects are launched and will be continuously monitored to provide specific assistance in complying with FAIR principles. Another issue is how to ensure that the achievements and results are best used within the NFDI4Earth community. For the next cohort, a strategy will be developed to increase the visibility of the work already in early stages, and to identify early on potential building blocks to be integrated into the NFDI community and

services. This strategy could be jointly developed under Task Area 1, as the other measures have similar issues.