NFDI4Earth

Milestone MS1.2.2

Calls for proposals, evaluation and selection

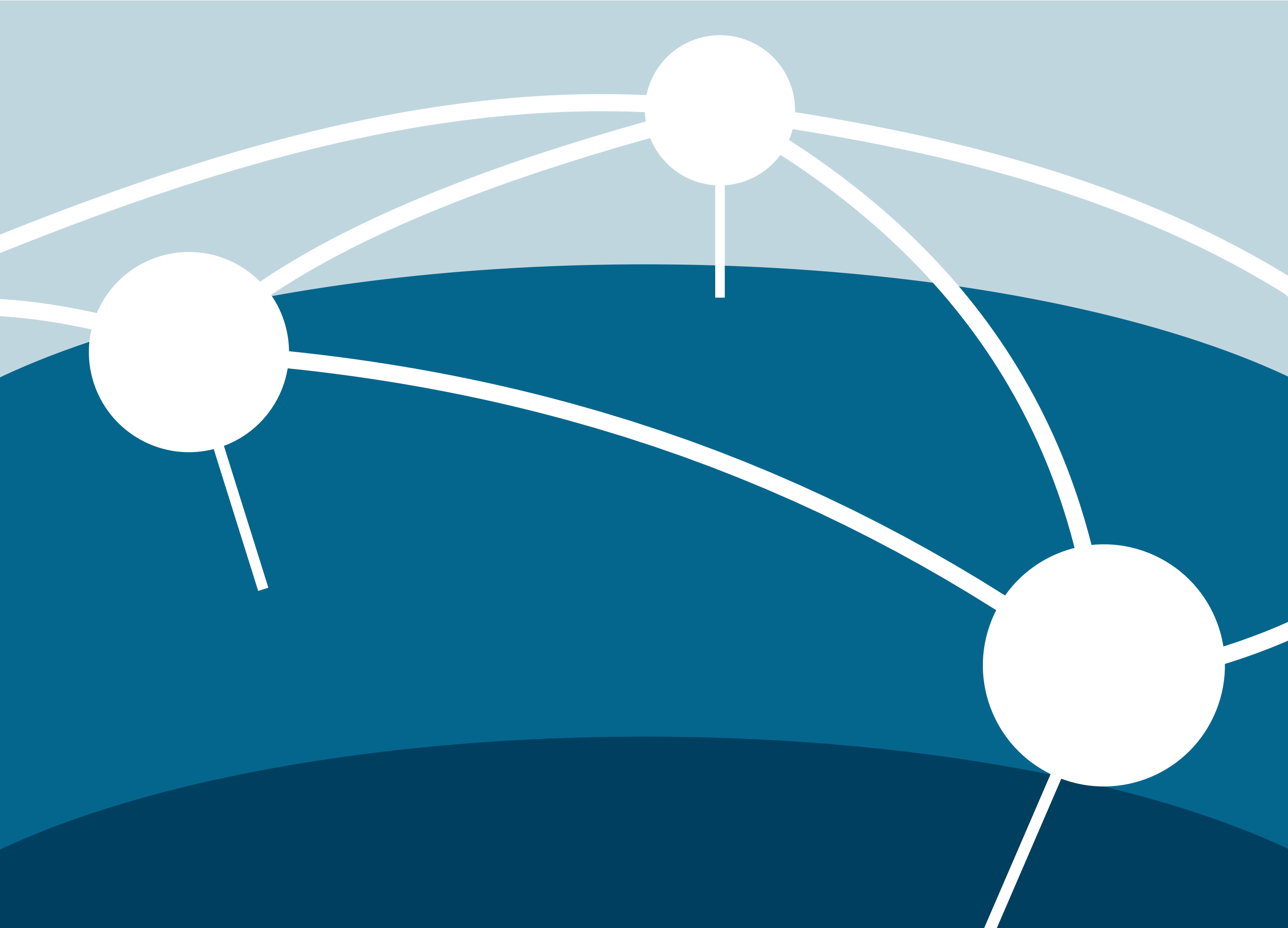
NFDI4Earth Incubators

Steffen Busch (steffen.busch@ikg.uni-hannover.de), Monika Sester, Hao Lie, Artem Leichter, Marco Kulüke, Ankita Ravi Vaswani

2022-09

DOI: 10.5281/zenodo.13811400

nfdi4earth.de



Citation

Steffen Busch, Monika Sester, Hao Lie, Artem Leichter, Marco Kulüke, Ankita Ravi Vaswani. 2024. *Calls for proposals, evaluation and selection (NFDI4Earth Milestone MS1.2.2) (NFDI4Earth Deliverable MS1.2.2)*. NFDI4Earth Community on Zenodo. https://doi.org/10.5281/zenodo.13811400

License

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 describes the results of the first call for proposals, provides statistics, and presents the first selected incubator projects with links to their results. The report is organized as follows: Section 1 refers too the April 2022 public call and presents its timeline, section 2 shows statistics of all submissions, and section 3 lists all selected proposals with their repositories, DOIs, and collaborators. Finally,in the appendix the public call and all original submissions of the accepted project proposals are presented.

Contributions

Based on CRedit Contributor Roles, see https://credit.niso.org/.

SB: Writing – original draft

MS: Writing – original draft

HL: Project administration

AL: Project administration

MK: Project administration

ARV: Project administration

Contents

1. First Call and Timeline

The first call sec. A has been published on 1.4.2022.

In order to allow for a quick implementation of new ideas, the whole process from start of the call till the decision was planned to last only approx. 2 months.

The following timeline for the first call was planned:

- Announcement: 01.04.2022
- Deadline for submission: 29.04.2022 (after 4 weeks)
- Reviewing process: 13.05.2022 (another 2 weeks)
- Decision: around 03.06.2022 (approx. 8-9 weeks after the announcement)
- Official start: mid July, August, September of 2022

2. Statistics

A total of 24 proposals has been submitted with topics from (areas as defined in the DFG disciplinary scheme):

- Atmospheric science, Oceanography and climate research (9 proposals)
- Geography (4 proposals)
- Geology and Paleontology (3 proposals)
- Mineralogy, Petrology and Geochemistry (3 proposals)
- Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartography (3 proposals)
- Biodiversity (1 proposal)
- Visualization (1 proposal)

In total, the projects required a funding of 17x 6 months, 3x 5 months, and 1x 3 months.

As for the institutions of the applicants, there were 3x - Hereon, 3x DLR, 3x Leibniz University Hannover, 2x University Bremen, 1x DKRZ, FZ-Jülich, IGB-Berlin, ...

fig. 1 give a more detailed view of the applicants and their institutions.

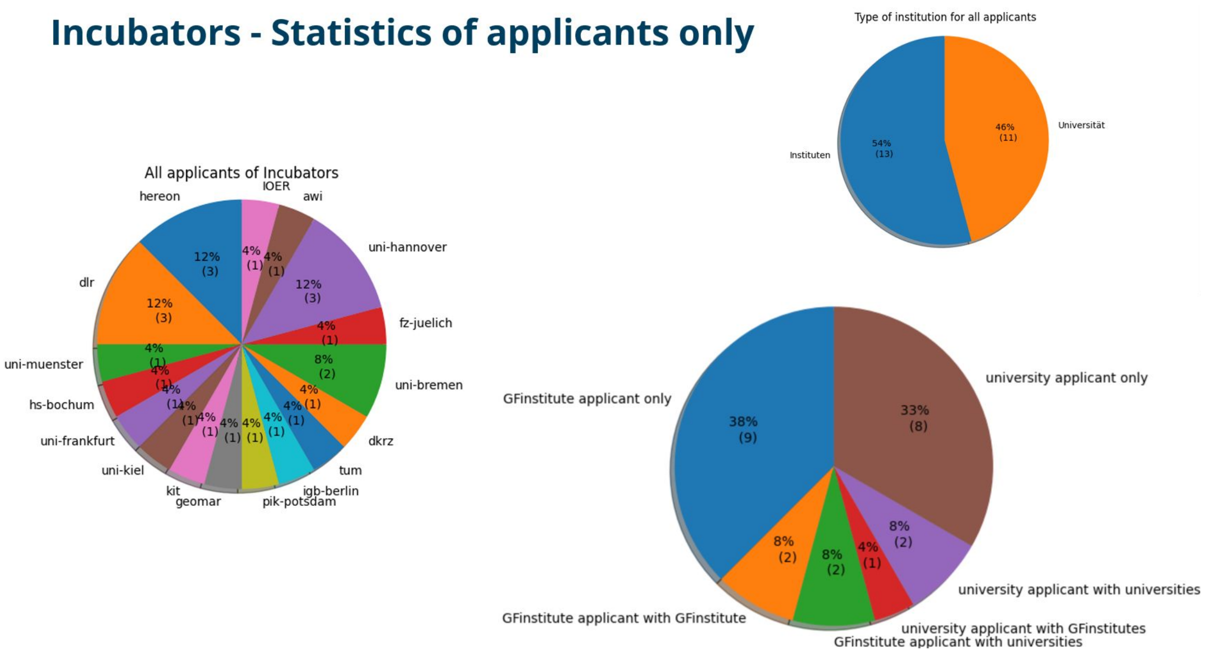


Figure 1: Overview of institutions and applicants

3. Selected Proposals - Overview

The available funding allowed to grant the following four proposals (the proposals can be found in the appendix; their deliverables are available via the respective links given below):

- IPFS Pinning Service for Open Climate Research Data sec. B

  Author: Marco Kulüke

  Orcid: https://orcid.org/0000-0003-0611-2567

  Affiliation: DKRZ

  Deliverables: GitLab and DOI:/10.5281/zenodo.7646356

- scrAiber: Data Mining Driven Microscopic Reference Data Acquisition

  Author: Artem Leichter

  Orcid: https://orcid.org/0000-0002-3524-2216

  Affiliation: Institute of Cartography and Geoinformatics, Leibniz Universit ̈at Hannover, Hanover

  Deliverables: GitLab and DOI:10.5281/zenodo.7744225

- New framework for analysis of aquatic ecosystems sec. D

  Author: Ankita Ravi Vaswani

Orcid: https://orcid.org/0000-0001-5015-8525

Affiliation: Biological Carbon Pump, Helmholtz Center Hereon

Deliverables: GitLab and DOI:10.5281/zenodo.7763865

- Hierarchical Data Format for Water-related Big Geodata (HDF4Water) sec. E

  Author: Hao Lie

  Orcid: https://orcid.org/0000-0002-6336-8772

  Affiliation: Technische Universit □at M □unchen

  Deliverables: GitLab and DOI:10.5281/zenodo.7562587

A. First Call

(included document begins on next page)

Preamble - Incubator Projects for the NFDI4Earth

1st call - Submission Deadline: May 13, 2022

The objective of the Incubator Lab is to foster the exploration of new, potentially relevant building blocks to be included in the NFDI4Earth and related NFDIs. The Incubator Lab invites proposals on novel cutting-edge tools, latest innovations, and high-risk blue-sky ideas. Incubator Projects are supposed to be small, concise, and have a clear focus. The funding duration of each project is therefore limited to 3-6 months for one researcher (DFG personnel rate). In total, in this call up to 8 projects can be funded.

Eventually, all the funded projects will contribute to a repository of advanced tools to improve the use and usability of NFDI4Earth data, advancing the state-of-the-art and to better meet (current and future) user needs.

How to apply for an NFDI4Earth Incubator Project?

Submissions should focus on tools and methods, and can relate to all aspects of the NFDI4Earth, including, but not limited to:

- automatic metadata extraction and annotations,
- machine learning, data interpretation, data fusion,
- visualization and interaction,
- …

Expected outputs should be, for example

- reproducibility software,
- easy to install and well documented with example data (preferably in GitLab),
- easy to apply, including small tutorials or examples, e.g., as a Jupyter-notebook,
- …

Application and evaluation process:

Any researcher can apply, who is affiliated or hosted by a German research institution. In the latter case, a letter of confirmation from the host is required along with the proposal submission. (If you do not have a host, please feel free to contact the NFDI4Earth Coordination Office.)

Incubator Projects must be described concisely; the proposal must not exceed 3 pages. Submissions should follow the template (Template Incubator). Proposals for Incubator Projects will be evaluated in an open review process. Both the proposals and the anonymous reviews will be made accessible. By submitting a proposal, you as PI agree to an open-access evaluation of the proposal.

The proposal will be assessed in three categories – each category will be scored from 1 (poor) to 5 (excellent) and a total grade will be derived using the weighting indicated below:

- Quality and Innovation (weight 40%)
  - Is the idea convincing?

  - Does it solve an interesting / relevant problem?
  - Is it new (in the field)?
  - Is the ambition adequate?
- Feasibility (weight 40%)
  - Does the researcher have the required expertise?
  - Is the research plan feasible?
  - Does it use or extend existing technologies? Or is it a novel innovation?
- Potential for transferability (weight 20%)
  - Ability to process similar data from different time periods / locations / institutions

We want to be maximally inclusive and invite new partners to the NFDI4Earth. Therefore, given equal quality of proposals, in selection process

- proposals from new partners are preferred,
- shorter projects are preferred.

Please submit your proposal via:

https://www.geo-x.net/nfdi4earth-incubators/application-form/

Note: If the submitted incubator project proposal is successful, the applicant and host (if applicable) agree that

- the PIs parent institution will be named as Participant (with the PI as representative) of the NFDI4Earth,
- a letter of commitment (LoC) will be provided if the institution of the PI is not yet a Participant in NFDI4Earth (due to DFG regulations the letter has to be signed by your institution),
- the so called ‘funds transfer and cooperation agreement’ must be signed as Participant (Institution) for funding.

Also note: The submission is subject to a review process and does not guarantee any funding!

FAQ Incubator Projects

What should be included in the LoC?

- Running an NFDI4Earth Incubator Project requires you to become a Participant in NFDI4Earth. If your institution is not yet a Participant in NFDI4Earth, we would ask you to send a Letter of Commitment. The template for the partner’s LoC is available on request.

For what kind of resources can we apply?

- You may only apply for staff costs, 3 to 6 months full-time equivalent (FTE).

I have a great idea, but I am not affiliated to a research organization. Can I still participate?

- Yes, good ideas are welcome! Please find a host either from the list of participants, or from a German research organization. Your host has to confirm the willingness. If you do not find a host, please contact us.

From whom can I request further information or answers to further questions?

- Feel free to send an email to nfdi4earth-info@groups.tu-dresden.de (Coordination Office) or yu.feng@ikg.uni-hannover.de

B. IPFS

(included document begins on next page)

IPFS Pinning Service for Open Climate Research Data

Marco Kulüke*, Stephan Kindermann*, Tobias Kölling**

* Deutsches Klimarechenzentrum
**Max-Planck Institute for Meteorology

Date of submission: 13.05.2022

Required funding: 4 months for one researcher at DKRZ

Abstract

Making data FAIR requires not only trusted repositories but also trusted workflows between data providers and infrastructure providers. Limited data access, unintentional and unnoticed data changes or even (overlooked) data loss pose great challenges to those involved. This incubator project aims to mitigate these challenges by exploring an easy-to-use data management service for researchers based on the InterPlanetary File System (IPFS), an emerging distributed web technology, which ensures data authenticity and fault-tolerant remote access. Based on a transferable prototypical implementation to be built within the DKRZ infrastructure, the suitability of the IPFS for a distributed and secure "web" for research data is being examined.

I. *Introduction*

Reliable and secure data exchange among scientists and between infrastructure providers is essential to enable FAIR data workflows. A major problem with research projects is that the infrastructure provider only receives and takes over responsibility for the final data at the very end of the project and has little ability to guide the data management process during the project. Consequences of insufficient data management can be unrecognized subsequent data changes or even data loss.

Recently there has been a growing interest in using the Interplanetary File System (IPFS)[1] to conquer this problem. The IPFS has an active community and has been developed since 2015 as an open-source peer-to-peer storage network for sharing data in a distributed file system. It is based on *Content Addressable Storage*. Unlike traditional location addressable storage, content addressable storage ensures that data is immutable by assigning a cryptographic hash to each block of data. Due to its distributed nature, the IPFS is also fault-tolerant and ensures workflows even if individual infrastructure components fail. Moreover, its public file sharing architecture strengthens Open Science efforts. Previous projects have shown that the IPFS can act as a repository for field experiment data. For example, this was demonstrated by storing data from the EUREC4A[2] field campaign on IPFS. However, there is still little experience with climate model

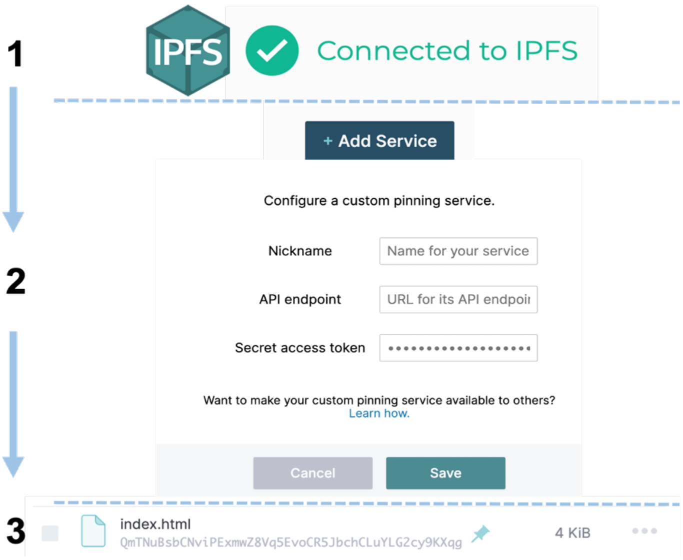
[1] https://ipfs.io/
[2] https://eurec4a.eu/

data management on IPFS. Therefore, this study examines the possibility of storing climate model datasets on IPFS.

For DKRZ, this is particularly interesting because it offers the opportunity to better guide scientists through the data management process during the project phase and reduce data inconsistencies and data loss.

II. *Incubator Project description*

Given the short duration of this project, this incubator will focus on only a few aspects. The aim of the incubator project is to provide researchers with a *Pinning Service*. Pinning is an important concept in IPFS, because it tells IPFS to always keep an object. The proposed outcome of the project will be a so-called *third-party remote pinning service*, which will be a simple-to-use API endpoint at an infrastructure provider such as DKRZ, which allows researchers to store a copy of their data.



Configure a custom pinning service.

Want to make your custom pinning service available to others? Learn how.

index.html
QmTNuBsbCNviPExmwZ8Vq5EvoCR5JbchCLuYLG2cy9KXqg

Figure 1 Visualization of the proposed workflow for researchers. (pictures taken from IPFS Desktop App)

In the planned workflow, as shown in figure 1, the researchers download the IPFS in the first step and install[3] it on their device. In the second step, they add the infrastructure provider's pinning service API endpoint. These steps only need to be applied once on a given device. Finally, the researchers simply add a pin to the data to be secured. In this way, the data is available anytime, everywhere and thus serves as a backup that remains even after the original data has been deleted by the researchers. Adding to this point is the native IPFS benefit of immutable data.

The suggested timeframe for this project is 4 months. The **initial preparation phase** is to understand how to set up the IPFS pinning service. This phase also includes an exchange with existing users of the IPFS and especially with the providers of the IPFS infrastructure for the EUREC4A campaign. The following **technical implementation phase** includes setting up the IPFS pinning service on a virtual machine within the DKRZ infrastructure. The subsequent **test phase** includes the simulation of a prototype project. Furthermore, a comprehensive **report** detailing our experiences of the use of the IPFS for climate data management with an assessment for potential future applications, a user guide, a documented GitLab repository, and explanatory

[3] https://docs.ipfs.io/install/

Jupyter notebooks will be produced. In addition, participation in the still young community is envisaged.

The DKRZ includes both the infrastructure and the technical know-how to implement this project. In all phases of the project, the quality of the proposed service is evaluated for example by conducting tests with different file sizes and corrupted files or by looking at the effect of file alterations. All tests will be performed with climate model data, which are typically large multidimensional datasets. This will generate new experiences with chunked data on IPFS. Docker containerization will ensure the portability of the developed software and will help other infrastructure providers to provide a similar service for their users.

III. *Relevance for the NFDI4Earth*

The proposed project fits perfectly with NFDI4Earth's goal of "providing simple, efficient, open and unrestricted access to all relevant Earth system data"[4]. Due to its use of the cutting-edge technology of the IPFS and its few existing applications in the field of climate science, the project has an innovative character and the potential to open new opportunities. Researchers and infrastructure providers benefit equally from the project. Already during the project phase, the researchers receive an easy-to-use tool for data exchange. The Pinning Service acts as a backup, prevents data loss and ensures data authenticity. On the part of the infrastructure provider, the tool enables insights into the data management already during the project phase and thus reduces duplication of work.

The prototype of the incubator project will be realized with data from the DKRZ CMIP Data Pool[5], which is part of the NFDI4Earth repositories and infrastructures.

IV. *Deliverables*

The project outcomes will be delivered in the following structure:

- Comprehensive report with experiences of the use of the IPFS for climate data management and an assessment for potential future applications
- Guide for users and infrastructure providers
- Open-source repository on GitLab[6]
- Tutorial Jupyter Notebooks
- Repository Documentation with workflow description
- Docker Image with all relevant software for infrastructure provider

[4] https://www.nfdi4earth.de/about-us

[5] https://cmip-data-pool.dkrz.de/

[6] https://gitlab.dkrz.de/data-infrastructure-services/ipfs-pinning-service-for-open-climate-research-data

C. scrAiber

(included document begins on next page)

NFDI4Earth

scrAiber:

Data Mining Driven Microscopic Reference Data Acquisition

M. Sc. Artem Leichter

Institute of Cartography and Geoinformatics
Leibniz University Hannover

Dr. Renat Almeev
Prof. Dr. rer. nat. Francois Holtz

Institut of Mineralogy
Leibniz University Hannover

Date of submission 13.05.2022

Required funding (**5 months**)

Abstract

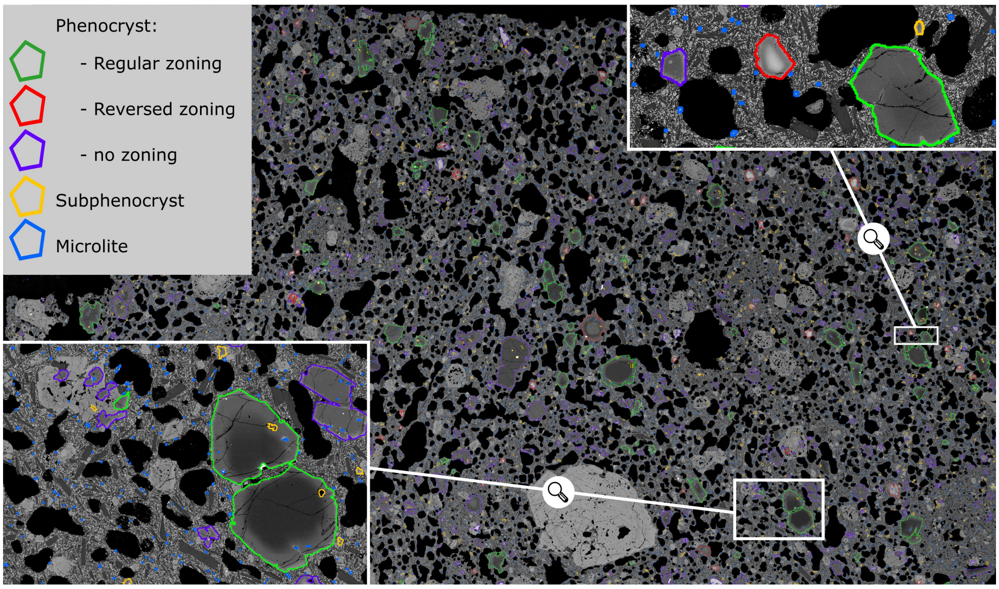
Creating training datasets for machine learning (ML) applications is always time consuming and costly. In domains where a high degree of expertise is required to generate the reference data, the corresponding costs are high and thus slow down the use of artificial intelligence (AI) systems. This proposal focusses on automated mineralogy and will provide tools to characterize the microscopic textural and mineralogical features of thin sections of rocks using back scattered electron images. Our goal is to address this problem with a data mining application where unsupervised methods in combination with expert users generate reference data without additional effort and cost for explicit labeling. The tools will be developed so that it can be used by scientists that have not a profound knowledge of ML.

I. *Introduction*

Availability of data and reference data is crucial for the success of ML algorithms. Creating reference data is costly and time consuming, especially for the characterization of rock samples in Earth sciences. The mineralogy and textures of rock samples can be characterized by a variety of analytical techniques, and detailed investigations are mostly based on the microscopic analysis of thin sections. The most common, fast and high resolution technique used to characterize thin sections is based on the acquisition of back scattered electron (BSE) images using electron microscopy. Such images can be used to extract information for an extremely large range of applications in mineralogy. However, the quantitative analysis of rough data requires experts, whose time is costly and limited. To address this problem, we propose a workflow combining unsupervised ML methods with expert input. Our prototype implementation of the approach will focus on the (unsupervised) *segmentation* of BSE images acquired on modern scanning electron microscopes with subsequent *clustering and classification* of objects of interest (e.g. minerals).
Unsupervised machine learning (ML) methods extract information without the need for reference data, although it is always advantageous to have reference data to validate the models or at least identify the suitable metaparameter. For example, a common strategy for creating reference data is to use unsupervised ML methods in the first step and then to refine the results through user

interaction. Such approach requires high expertise of the user. We propose to prototype a digital environment that provides productive work based on unsupervised ML methods so that users interact with it on their own initiative while working on their own tasks. In this way, there is no additional workforce investment necessary. The reference data is created implicitly by tracking the user interaction. This approach is state of the art in many commercial applications such as online stores and video platforms, but as far as we know it is novel in the context of digital petrological analysis.

From our preliminary work based on the analysis of BSE image data, we have a number of implemented algorithms that yielded extremely promising results but that can only be carried out by users with strong programming skills (Leichter et al., 2022). Based on our experience gained so far and in the context of this project, we plan to construct a platform which can be used for image analysis by non-experienced user.



1. Examples of successful deep learning (DL) segmentation and rule based classification applied on a thin section of a volcanic rock (Leichter et al., 2022). In this example, ML was used to identify automatically all olivine crystals present in the thin section (more than 20.000) and to classify the minerals depending on the compositional zoning as well as the type of minerals (phenocrysts, subphenocrysts, microlites). Phenocrysts with compositional zoning (ca. 800 crystals) were subsequently used to apply diffusion chronometry. Without application of ML learning techniques, it is impossible to treat manually (and objectively) this amount of data. For more examples visit icaml.org/olmap/.

II. *Incubator Project description*

This project consists of two main packages, namely (1) implementation of a productive online tool for analyzing BSE data and (2) acquisition of user tracking data and its prototypic evaluation.

(1) Implementation of a productive online tool (3 Month): The online tool, hereafter referred to as "scrAiber", is crucial for the possibility of collecting data on segmentation. For this purpose, scrAiber should allow the user to solve specific problems (e.g. semantic segmentation of minerals, clustering of crystals, unsupervised pixelwise segmentation) when working with BSE images. In order to meet the compact framework of this project, the system is implemented as a web application, which facilitates the management of dependencies and increases availability for users. The implemented functionality is based on already existing solutions like preprocessing of BSE raw data, segmentation of minerals with deep learning (DL) framework or unsupervised segmentation. The user interface (UI) is implemented in a minimalistic way with focus on functionality and a low entry barrier (immediate start of work).

(2) Collecting user tracking data (2 Month): User tracking data allows applications such as online stores to identify similar and complementary items based on user activity. Actions such as completed purchase of a product and joint purchase of multiple products serve as indicators for the analysis. In this project, such indicators must first be identified and integrated into the scrAiber workflow. In this step, mineralogy expert(s) actively work with scrAiber. The outcome of their work session is documented in short interviews and questionnaires. The gathered data is used to prototypical evaluation of the tracking data.

For the processing of the two packages, partner institutions bringing experts in Earth Sciences (mainly petrology/mineralogy) and geoinformatics (data engineering competence) together are required. The processing time of the first package is estimated to be three months and consists of two man-months work for an expert in Data Engineering (Institute of Cartography and Geoinformatics) and one month for a petrologist (Institut of Mineralogie) . For the second package, the working time is composed of one man-month work for each partner.

III. *Relevance for the NFDI4Earth*

Machine learning opens up unique opportunities for the analysis of *electron microscopy image data* (this proposal) but also for a variety of analytical sensors in Earth sciences in general. To be able to use these possibilities, remote data are mandatory. The strategy proposed here can also be applied to other domains to acquire reference data with little additional effort.

The main users of scrAiber are expected to be petrologists investigating and characterizing rock systems. The application of scrAiber is expected to be helpful in applied geosciences and for mineral processing engieneers, allowing a fast characterization of a series of rocks such as drill cores (e.g. drill cores used for the characterization of ore deposits) as well as for other more fundamental scientific projects based on mineralogical analyses (e.g., structural geology, volcanology, metamorphic and magmatic rocks). The tool will also be extremely useful for teaching courses in Earth Sciences and can also be used to introduce the advantages of ML.

So far, to our knowledge, there is no structured open access database available for BSE images of thin sections. Thus, at this stage, our proposal will not establish a tool that can be directly applied to available databases. However, the tool developed in this proposal, allowing users to extract extremely quickly a huge amount of information from BSE images, is expected to initiate and promote the creation of databases for the microscopic analysis of thin sections.

The main repositories that can be used are rocks/thin sections from drill cores or field expeditions collected in the frame of several national research initiatives or international expeditions. These rocks are stored in various places in Germany. The access to data and to rocks from national drilling cores, stored at repositories under the responsibility of the Bundesanstalt für Geowissenschaften and Rohstoffe (BGR) is possible in the frame of the cooperation of the Leibniz University of Hannover and the BGR. Cores and thin sections that have been characterized in detail can also be obtained by contacting the ICDP and IODP coordinators in Germany (GFZ Potsdam and BGR, respectively). The petrologists of the Institut of Mineralogy have also numerous cooperations which would allow a rapid access to numerous samples and BSE data from thin sections.

IV. *Deliverables*

1. scrAiber: A tool for segmenting SEM data that is accessible to all via the internet.
2. scrAiber Code: opensource project for the scrAiber online tool, wich allows the adoption to different datatyper and tasks.
3.The data gathered by scrAiber will be published as open data and can be integrated to NFDI4Earth repositories.

V. *References*

Leichter, A., Almeev, R. R., Wittich, D., Beckmann, P., Rottensteiner, F., Holtz, F., & Sester, M. (2022). Automated Segmentation of Olivine Phenocrysts in a Volcanic Rock Thin Section Using a Fully Convolutional Neural Network. Front. Earth Sci, 10, 740638. | https://doi.org/10.3389/feart.2022.740638

D. Aquatic Ecosystems

(included document begins on next page)

NFDI4Earth

New framework for analysis of aquatic ecosystems

Ankita Ravi Vaswani and Klas Ove Möller[1]

[1] Biological carbon pump group, Institute of Carbon Cycles, Helmholtz-Zentrum Hereon, Geesthacht

Date of submission: 13.05.2022
Required funding: 6 months

Abstract

Advances in high-throughput *in situ* imaging offer unprecedented insights into aquatic ecosystems by observing organisms in their natural habitats. However, unlocking this potential requires new analysis tools that transcend species identification to reveal morphological, behavioral, physiological and life-history traits. We will develop, document and validate an image analysis pipeline for semi-automated functional trait annotation, apply it to zooplankton in a continuously monitored North Sea region, and train a neural network for full automation. We foresee that these tools will enable new avenues of investigation in aquatic research, ecosystem modelling and global biogeochemical flux estimations, revealing previously inaccessible relationships between species biodiversity, zooplankton traits and seasonal variations in environmental conditions.

Introduction

Zooplankton are essential for aquatic food webs and biogeochemical cycles[1,2]. Understanding how changing environmental conditions affect distribution and abundance of zooplankton traits allows us to decode the effects of climate change on biodiversity, ecosystem dynamics and global carbon cycles[3–5]. Body size is a master trait linked to developmental stage, respiration, metabolism, excretion, motility and predator-prey interactions[6]. *In situ* images also provide information about individual organisms' survival, growth, reproduction and resource acquisition from the visual signatures of the underlying traits[4,7] such as lipid reserves[8], egg clutch sizes[9], appendage extension[7], and body posture[7]. Recently, zooplankton traits analyzed during spring ice-melts in the Arctic Ocean have revealed complex ecosystem responses to environmental changes[7]. However, manual annotation procedures cannot scale to analyze the billions of images now produced by imaging technologies across diverse environmental conditions and long durations. Instead, we require new analysis pipelines to efficiently annotate a fraction of these data, to produce training data for fully automated deep-learning based feature extraction[5,10].

We will construct an image analysis pipeline (deliverable **D1**), combining an existing neural network for taxonomic classification[11] with semi-automated segmentation of zooplankton traits. Such a consolidated pipeline does not currently exist for trait extraction and will be an invaluable tool for marine researchers. We will apply D1 to a continuously monitored North Sea region to produce an annotated trait dataset (**D2**), and use this to train a convolutional neural network (CNN, **D3**) in collaboration with Hereon's Model-Driven Machine Learning group (MDML). We will document each step of this process using interactive tutorials (**D4**) and publish open-source software (**D5**). These deliverables will accelerate trait extraction and increase reproducibility.

During my PhD, I designed a semi-automated analysis pipeline to extract morphological features from ~15,000 embryonic brain cells, revealing complex relationships between cell morphology and migration speed[12]. Despite the new application domain, the technical aspects and challenges of semi-automated segmentation align well with my skill set well, allowing me to design an efficient feature extraction pipeline for the annotation of zooplankton traits.

I. *Incubator Project description*

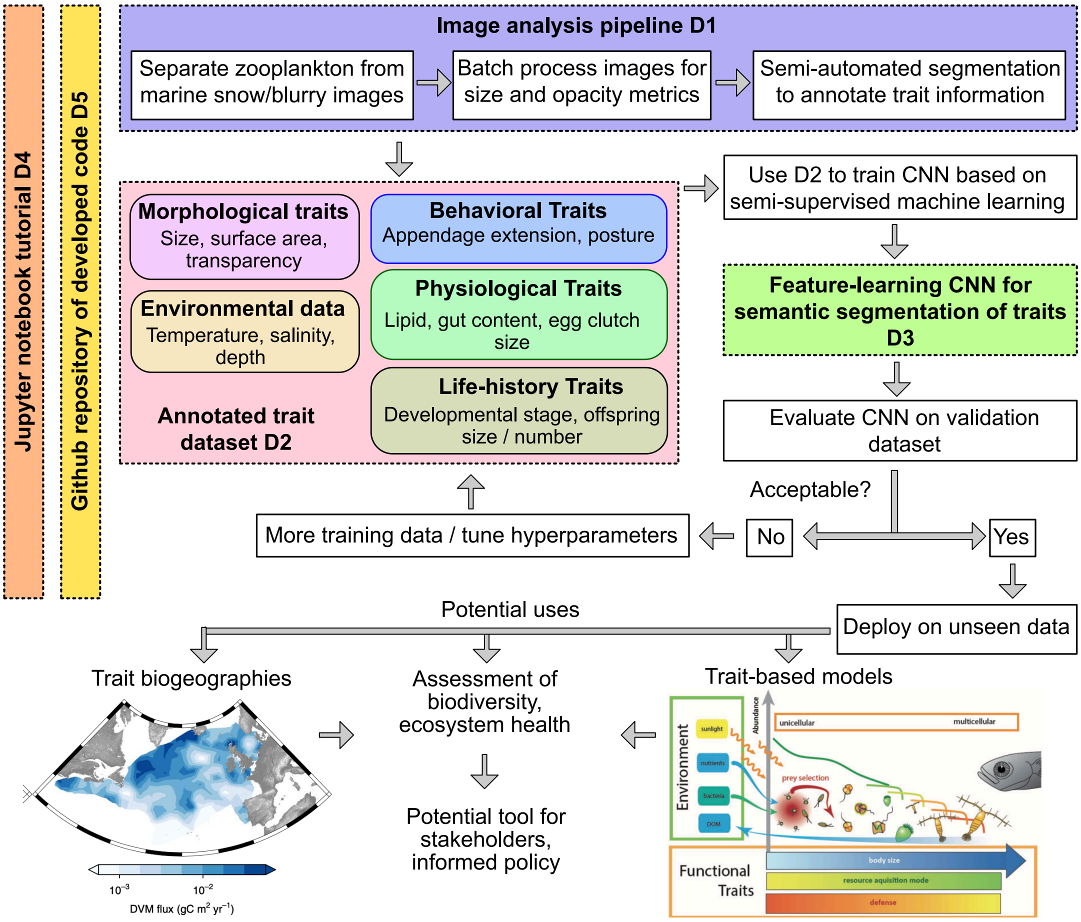

Jupyter notebook tutorial D4
Github repository of developed code D5
Image analysis pipeline D1
Separate zooplankton from marine snow/blurry images
Batch process images for size and opacity metrics
Semi-automated segmentation to annotate trait information
Morphological traits
Size, surface area, transparency
Behavioral Traits
Appendage extension, posture
Environmental data
Temperature, salinity, depth
Physiological Traits
Lipid, gut content, egg clutch size
Life-history Traits
Developmental stage, offspring size / number
Annotated trait dataset D2
Use D2 to train CNN based on semi-supervised machine learning
Feature-learning CNN for semantic segmentation of traits D3
Evaluate CNN on validation dataset
Acceptable?
More training data / tune hyperparameters
No
Yes
Deploy on unseen data
Potential uses
Trait biogeographies
10^{-3}
10^{-2}
DVM flux (gC m^2 yr^{-1})
Assessment of biodiversity, ecosystem health
Potential tool for stakeholders, informed policy
Trait-based models
Environment
sunlight
nutrients
bacteria
DOM
Abundance
unicellular
multicellular
prey selection
Functional Traits
body size
resource aquisition mode
defense

Figure 1: A consolidated data analysis pipeline D1 will be built to classify *in situ* plankton images, extract morphological features and guide users through semi-automated trait segmentation. D1 will be applied to HPT data to generate annotations D2. D2 will be used to train a trait extraction CNN (D3). A Jupyter notebook (D4) will guide users through customizing and implementing D1 and D3 for diverse image data. Wide-scale application of these tools will lead to detailed trait-biogeographies, better ecosystem models and accurate assessment of the impact of climate change on biodiversity and ecosystem health. Image adapted from previous studies[13,16].

The Möller lab has deployed underwater observatories in the North Sea to acquire high resolution plankton time series (HPT) imaged at 6-10 Hz. ~10% of HPT data will be analyzed using our analysis pipeline (D1) to generate an annotated trait dataset (D2).

1. In the first 2 weeks, zooplankton will be separated from marine snow and blurry images and classified into relevant taxon units (copepods, jellyfish, marine snow, etc.) using an existing CNN developed by the Möller and MDML groups[11].
2. Quantitative traits such as body length, surface area, volume and opacity metrics will be automatically extracted from images[13]. Taxonomic classification will be confirmed and semi-automated segmentation will be used to extract complex traits such as lipid body area, lipid fraction[8], egg-clutch size[9], body posture and extension of appendages and antennae[7].
3. D1 will combine all automated and semi-automated methods from steps 1-2 into one program, where imported images will be sequentially classified, processed to extract numerical descriptors and inspected by a user for trait segmentation. All measurements and trait segmentation masks will be stored as metadata linked to the images. This pipeline will run in a browser with a Jupyter-based rich graphical interface (D4). Development of the image analysis pipeline (D1) will require ~2-3 months, culminating in an annotated trait dataset (D2).
4. D2 will be used for semi-supervised training of a new CNN (distinct from the existing CNN of step 1), resulting in software for automatic trait extraction (D3). This step will be carried out over ~2-3 months in collaboration with the MDML group and support from Helmholtz AI consultants (see attached letters of support). If further annotated data are needed, step 3 will be carried out on additional images from the HPT data in parallel to CNN training.
5. Tutorials for the the data analysis pipeline and analyses in steps 1-3 will be documented in Jupyter notebooks (D4). Code will be documented and shared in open-source repositories (Github; D5). This step will be carried out in parallel to previous steps.

This project has a high chance of success given the participants' expertise and experience and immediate availability of required tools. D1-2 will provide value to end users prior to the completion of subsequent steps. Beyond the scope of this incubator project, follow-up work is foreseen investigating seasonal variations in trait abundances in the North Sea using D2 and D3.

II. *Relevance for the NFDI4Earth*

- Automatic taxonomic classification and trait extraction (D1/D3) will be valuable for marine biologists, ecologists and image analysts.
- Global trait biogeographies, generated using tools developed here, could reveal relationships between trait abundances, biodiversity and ecosystem health, helping scientists, stakeholders, and policy makers to guide sustainable ecosystem management[14,15].
- Detailed trait-biogeographies could also be overlayed on data from NFDI4Earth repositories (e.g., HZG marine geoportal coastal maps) to reveal relationships between the biology and geography of aquatic ecosystems.
- Our tools could increase incorporation of functional trait data in detailed, predictive models of aquatic ecosystems and carbon flux[16,17].

III. *Deliverables*

- D1: Data-analysis pipeline (month 3)
- D2: Annotated trait dataset (month 3)
- D3: CNN for automated feature learning and extraction (month 6)

- D4: Tutorials for design and implementation of D1 (month 3), generation of D2 (month 3) and D3 (month 6)
- D5: Open-source code repository (GitHub)

References

1. Frederiksen, M., Edwards, M., Richardson, A. J., Halliday, N. C. & Wanless, S. From plankton to top predators: bottom-up control of a marine food web across four trophic levels. *J Anim Ecology* **75**, 1259–1268 (2006).
2. Boyd, P. W. Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature* **9** (2016).
3. Violle, C. *et al.* Let the concept of trait be functional! *Oikos* **116**, 882–892 (2007).
4. Ohman, M. D. A sea of tentacles: optically discernible traits resolved from planktonic organisms in situ. *ICES Journal of Marine Science* **76**, 1959–1972 (2019).
5. Orenstein, E. *et al.* Machine learning techniques to characterize functional traits of plankton from image data. *In press* (2022).
6. Hatton, Heneghan, Bar-On & Galbraith. The global ocean size-spectrum from bacteria to whales. *Sci. Adv.* **7**, (2021).
7. Vilgrain, L. *et al.* Trait-based approach using in situ copepod images reveals contrasting ecological patterns across an Arctic ice melt zone. *Limnol Oceanogr.* **66**, 1155–1167 (2021).
8. Schmid, M. S., Maps, F. & Fortier, L. Lipid load triggers migration to diapause in Arctic Calanus copepods—insights from underwater imaging. *Journal of Plankton Research* **40**, 311–325 (2018).
9. Möller, K. O. *et al.* Effects of climate-induced habitat changes on a key zooplankton species. *Journal of Plankton Research* **37**, 530–541 (2015).
10. Irisson, J.-O., Ayata, S.-D., Lindsay, D. J., Karp-Boss, L. & Stemmann, L. Machine Learning for the Study of Plankton and Marine Snow from Images. *Annu. Rev. Mar. Sci.* **14**, 18.1–18.25 (2021).
11. Schanz, T., Möller, K. O., Ruehl, S. & Greenberg, D. Robust Detection of Marine Life with Label-free Image Feature Learning and Probability Calibration. *Helmholtz AI Conference, Dresden* (2022).
12. Vaswani, A. R. *et al.* Correct setup of the substantia nigra requires Reelin-mediated fast, laterally-directed migration of dopaminergic neurons. *eLife* **8**, e41623 (2019).
13. Kiørboe, T., Visser, A. & Andersen, K. H. A trait-based approach to ocean ecology. *ICES Journal of Marine Science* **75**, 1849–1863 (2018).
14. Gorsky, G. *et al.* Digital zooplankton image analysis using the ZooScan integrated system. *Journal of Plankton Research* **32**, 285–303 (2010).
15. Prowe, A. E. F., Visser, A. W., Andersen, K. H., Chiba, S. & Kiørboe, T. Biogeography of zooplankton feeding strategy. *Limnol. Oceanogr.* **64**, 661–678 (2019).
16. Martini, S. *et al.* Functional trait-based approaches as a common framework for aquatic ecologists. *Limnol Oceanogr* **66**, 965–994 (2021).
17. Brun, P. *et al.* Climate change has altered zooplankton-fuelled carbon export in the North Atlantic. *Nat Ecol Evol* **3**, 416–423 (2019).
18. Serra-Pompei, C., Soudijn, F., Visser, A. W., Kiørboe, T. & Andersen, K. H. A general size- and trait-based model of plankton communities. *Progress in Oceanography* **189**, 102473 (2020).

Helmholtz-Zentrum
hereon

Helmholtz-Zentrum hereon | Max-Planck-Straße 1 | 21502 Geesthacht, DE

Geesthacht, 12.05.2022

Institut für
Dynamik der Küstenmeere
Mariner Schnee und Plankton
Institute of
Coastal Ocean Dynamics
Marine Snow and Plankton

www.hereon.de

Abteilungsleitung | Department Head
Dr. Klas Ove Möller
T +49 4152 87-2371
klas.moeller@hereon.de

Letter of support for the NFDI4Earth incubator project “New framework for analysis of aquatic ecosystems”

To whom it may concern,

This letter expresses my strong and full support for the NFDI4Earth Incubator project proposal “New framework for analysis of aquatic ecosystems” put forward by Dr. Ankita Vaswani. The proposal outlines an ambitious and timely approach which will provide new tools to examine biodiversity, ecosystem functioning and ecosystem health in the ocean.

The candidate will work in the "Biological Carbon Pump" group at the Institute of Carbon Cycles, Helmholtz Zentrum Hereon. The Biological Carbon Pump group has a strong background in exploring marine plankton dynamics, biological-physical interactions and the role of the marine environment in regulating global climate. At the Institute of Carbon Cycles, the candidate will find a wealth of complimentary expertise ranging over diverse topics including biology, oceanography, physics and ecology.

All required infrastructure, computing and data storage resources are available at Hereon including cluster access, large computation and storage resources, high bandwidth data transfers, support for code and software exchange via HiFis, and access to additional resources via HiCore.

The candidate is well qualified and experienced in processing and analysing large image datasets derived from optical instruments, which is the key technical skill required for the success of the project. With respect to this highly interdisciplinary project; the candidate will have support and collaboration from the Model-driven Machine Learning Group, at the Helmholtz-Zentrum Hereon, giving her a strong chance to successfully complete the project and achieve her career goals.

Klas Möller

……………………………………….

(Dr. Klas Ove Möller)

Helmholtz-Zentrum
hereon

Helmholtz-Zentrum Hereon | Max-Planck-Straße 1 | 21502 Geesthacht, DE

NFDI4Earth Coordination Office
Incubator Projects

Institut for Coastal Resarch
Modellbasiertes
Maschinelles Lernen
Institute of Coastal Systems
Analysis and Modeling
Model-Driven Machine Learning

www.hereon.de

Abteilungsleitung | Head of Department
Dr. David S. Greenberg
T +49 4152 87-2133
david.greenberg@hereon.de

Letter of Support for Project "New framework for analysis of aquatic ecosystem"
Geesthacht, 12 May 2021

Dear NFDI$_4$Earth Project Committee,

I'm writing to express strong support for the project "New framework for analysis of aquatic ecosystem." Since 2020 I have lead the Model-Driven Machine Learning research group at Helmholtz Centre Hereon, focused on developing and applying artificial intelligence and machine learning (AI/ML) to improve understanding and predictability both physical and biological aspects of coastal systems.

A major challenge in effectively applying AI/ML methods in the earth sciences is the wide gulf between the artificial, sanitized datasets used by the deep learning community and real field data. *In situ* observations of biological systems exhibit complexity, noise, nonstationarity and class imbalance far beyond AI/ML benchmarks, so that promising AI/ML methods often fail in the real world.

By annotating image datasets with labeled morphological features and functional traits, and by developing a pipeline for further annotation, this project would significantly aid and accelerate the development of AI/ML methods for studying marine organisms, with considerable benefits for both marine biology and deep learning.

I am therefore happy to pledge our research group's support in training deep learning models on these datasets. In particular, we can provide existing convolutional networks for detecting and classifying zooplankton image features, developed and trained in an ongoing collaboration with Dr. Klas Ove Möller at Hereon. We will provide access to multi-GPU compute nodes for training and deploying these networks.

Best regards,

David Greenberg

DKRZ
DEUTSCHES
KLIMARECHENZENTRUM

DKRZ GmbH · Bundesstr. 45a · D-20146 Hamburg

NFDI4Earth Coordination Office
Incubator Projects

Ust-IdNr.: DE118713406
Steuernummer: 17/413/00377

Deutsche Bank Hamburg
BLZ 200 700 00
Konto-Nr.: 367518800
IBAN: DE51 2007 0000 0367 5188 00
BIC: DEUTDEHHXXX
Telefon: 040 - 460094 -
Telefax: 040 - 460094 - 270

E-Mail: weigel@dkrz.de

Ansprechpartner: Tobias Weigel

Ihr Zeichen	Ihre Nachricht vom	Unser Zeichen	Durchwahl	Datum
			-	13. Mai 2022

Letter of Support for project “New framework for analysis of aquatic ecosystem”

Dear NFDI4Earth project committee,

the Helmholtz AI Cooperation Unit (Helmholtz AI) is set up as a hub for applied artificial intelligence and machine learning across the whole Helmholtz Association. It consists of six local units, representing the six research fields of the Helmholtz Association, each of which hosts research groups and an AI consultant team. The AI consultant teams’ mission is to enable and facilitate the use and implementation of AI/ML methods through short and mid-term collaborations of up to six months. Their expertise can be solicited through a Voucher System accessible at no cost to all researchers; decisions to engage in collaboration are based on capability and capacity.

As consultants focussing on “Earth and Environment” based at DKRZ we are happy to support the project “New framework for analysis of aquatic ecosystem” according to the procedures of the voucher support system. We recognize the feasibility of the proposed approach and see the potential of transfer to other application cases in the marine sciences, but also beyond. We have worked on Computer Vision problems including semantic segmentation in several prior support activities, and also set up efficient procedures for the practical training and tuning workflows, and would be happy to lend such additional expertise to the project.

Yours sincerely,

Tobias Weigel

Dr. Tobias Weigel

E. HDF4Water

(included document begins on next page)

Hierarchical Data Format for Water-related Big Geodata (HDF4Water)

Hao Li[1], Martin Werner[1]

[1]Chair of Big Geospatial Data Management, Department of Aerospace and Geodesy, Technical University of Munich

13 May 2022, required funding: 6 months

Abstract

Humans rely on clean water for their health, well-being, and various socio-economic activities. To ensure an accurate, up-to-date map of surface water bodies, the often heterogeneous big geodata (remote sensing, GIS, and climate data) must be jointly explored in an efficient and effective manner. In this context, a cross-platform and rock-solid data representation system is key to support advanced water-related research using cutting-edge data science technologies, like deep learning (DL) and high-performance computing (HPC). In this incubator project, we will develop a novel data representation system based on Hierarchical Data Format (HDF), which supports the integration of heterogeneous water-related big geodata and the training of state-of-the-art DL methods. The project will deliver high-quality technical guidelines together with an example water-related data repository based on HDF5 with the support of the BGD group in TUM, with which the NFDI4Earth will consistently benefit from this incubator project since the solution can serve as a blueprint for many other research fields facing the same big data challenge.

I. *Introduction*

Water plays a key role in human health, well-being, socio-economic activities, and Sustainable Development Goals (SDGs). In the past two years, the COVID-19 pandemic has demonstrated the substantial importance of hygiene rules, sanitation, and adequate access to clean water for reducing the spread of infectious diseases and preserving the public health of millions.

Recently, the development of deep learning (DL) techniques shows great potential to facilitate up-to-date and large-scale water-related research in Earth System Science (ESS), but also poses substantial challenges, especially when considering multi-sensor and multi-modal geospatial big data, which are all heterogeneous in their nature. Therefore, there is an unprecedented need for an efficient and flexible data representation and paradigm for both raster-based geodata (e.g., Multispectral satellite images), vector-based geodata (e.g., OpenStreetMap (OSM)), or even point-based geodata (e.g., LiDAR and Photogrammetry point clouds), while few satisfactory big data solutions exist so far.

In HDF4Water, we propose a concise and rock-solid data representation system based on the state-of-art Hierarchical Data Format 5 (HDF5), a multi-objects-based format originated from the National Center for Supercomputing Applications (NCSA). The key advantages of HDF5 mainly include: (i) support for heterogeneous data (e.g., any n-dimensional datasets); (ii) portable and easy-to-share, with no vendor or platform lock-in; (iii) cross-platform, from laptops to massively parallel systems; (iv) fast I/O allows for access time and storage space optimizations; (v) keep metadata with data, streamlining big data lifecycles and pipelines. Moreover, HDF5 can provide seamless support for modern DL libraries (e.g, TensorFlow, PyTorch, etc.) in addition to a self-describing file system of raster, vector, and point-based geodata in a way such that water-related DL models can be directly and efficiently implemented on top of this data representation with limited (if at all) loss of lineage.

HDF4Water aims to bridge between the current relational model of GIS (e.g., spatial types as extensions of spatial database management systems) and the high-performance computing (HPC) domain (e.g., DL, distributed computing) in the water-related ESS research, where immediate memory random access is of importance. Therefore, it is something significantly different from the HDF5-based NetCDF representation, which is optimized for raster data only. As a final result, showcases of storing heterogeneous big geodata for various water-related applications in HDF5 and ZARR (an open-source modern implementation similar to HDF5) format shall be a paradigm for many other ESS research fields, or even outside the NFDI4Earth community, facing the same big data challenge.

II. *Incubator Project description*

Recently, the emergence of heterogeneous big geodata has led to significant performance boosting in the field of water-related research, which facilitates a larger-scale and faster-speed of monitoring surface water than ever before. Global Surface Water Layer (GSWL) demonstrated the promising capability of integrating remote sensing data (e.g., 30m Landsat data) with multi-sensor auxiliary data (e.g., digital elevation models (DEM), glacier data, urban settlement data, etc) in global-scale water mapping. In the meantime, intensive research efforts are dedicated to developing supervised machine learning (ML) methods, especially using deep learning, for accurate surface water mapping, which achieved superior performance than traditional indices-based method. Moreover, early attempt of harvesting open-source GIS data from OSM as training data for DL models can well address the lack of sufficient training labels during supervised learning. However, this big geodata, all of which are heterogeneous and in diverse formats, request distinct converting and processing steps in order to support a joint analysis, which could be time- and effort-consuming in any aspect.

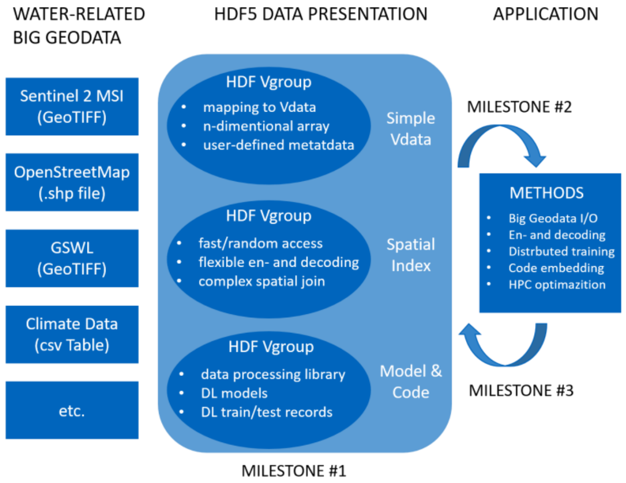


Figure 1: The data representation system based on HDF5

In the proposed Incubator project, we will develop a concise but robust data representation system (c.f. Figure 1) based on open-source technologies (HDF5 or ZARR) to facilitate DL applications with water-related big geodata. For this purpose, we draw on our preliminary work based on the Copernicus' Sentinel-2 MSI, OSM vector data, GSWL occurrence map in automatic national surface water mapping, and combine them with diverse climate data within the NFDI4Earth repositories to support a broader water-related ESS research in the future. To ensure the transferability, we rely on simplicity as a design factor and will integrate data-related source codes (e.g., scripts, Jupyter notebooks, etc.) and documents (e.g., PDF, markdown, HTML) right into the HDF5 data format bringing a full-fledged, but conceptually simple, data science container format to life. In principle, we want the whole data science project to be captured in a single file.

The HDF4Water project will run as follows: (i) first, by specifying the processing pipeline of water-related big geodata using GIS tool to feed an ImageDataGenerator for DL models trained for automatic water mapping in Germany; (ii) next, we design and implement a concise mapping of heterogeneous data into HDF5 Vgroups (c.f. Figure 1) providing a suitable random access and spatial index, where the source code of data en- and decoding will be embedded together with the raw data file to support future data exploitation in a fully reproducible way; (iii) last, we implement and showcase a distributed training regime (benchmarked on SuperMUC Next Generation) to specify how the source code and documents can be incorporated directly with the heterogeneous geodata. In addition, we plan to provide an example Jupyter-notebook with which this data representation can be decoded into individual files and encoded into an HDF container to enable interfacing with traditional GIS software.

III. *Relevance for the NFDI4Earth*

DL is currently one of the techniques that are more and more adopted in ESS research at large. However, traditional data formats (either raster or vector) are not optimized for the access patterns of deep learning from big geodata, therefore, stakeholders, from almost every stage of the data lifecycle, started to prepare datasets that are very suboptimal in terms of them being spatial datasets (e.g., fragmented, loss of metadata, etc.) or to compromise with comparably slow training and inference performance by relying on libraries from our field that have not been optimized for the workloads typical for DL. In this context, our incubator project aims to provide a single-technology solution for DL with heterogeneous big geodata based on modern HDF5 technology. The choice for HDF5, aside from its excellent features, is also based on the observation that HDF5 is an essential part of, e.g. TensorFlow, also available on most if not all DL libraries.

Though we select water-related big geodata as a case study, the HDF4Water can be regarded as a blueprint for similar research fields, and the lessons learned in this project contribute to the long-term vision of NFDI4Earth to provide researchers with FAIR, coherent, and open access to all relevant ESS data, to innovative big data management and data science methods.

IV. *Deliverables*

A. Technical guidelines and documents (Water-related big geodata in Germany)
   1. Format description for spatial data mapping into groups
   2. Metadata implementation and specification

B. An example data repository based on HDF5 and ZARR
   1. Raw water-related geodata including imagery and GIS data
   2. Complete source code (e.g., processing, DL models, distributed training)
   3. Markdown documentations and tutorials