FAIR@HPC – Improving HPC usage in ESS by FAIR data and compute services

a Concept Paper – v1.0 – May 2022
IG “High-performance Computing in Earth System Sciences” in NFDI4Earth

1 Motivation

There is a continuously increasing demand for using High-Performance-Computing (HPC) infrastructure for solving geoscientific questions in different domains. AI-based techniques and big data analytics are starting to play significant roles, while also the field of classical HPC simulations still experiences growth. As one prominent example we consider the weather and climate forecasting community which aims at improving prediction capabilities of simulation models by increasing resolution and multi-physics process descriptions. The amounts of input and output data handled in simulation, Big Data and AI have strongly grown, leading to challenges in effectively sharing data and making research reproducible. At the same time, the German research landscape is striving to make their data FAIR (Findable, Accessible, Interoperable, Reusable) within the NFDI initiative. However, none of the known data repositories can host PBytes of output produced in an HPC simulation. The pure transfer of input/output data becomes prohibitive for making data FAIR at typical HPC dataset sizes. Thus, it is mandatory to develop concepts for allowing a FAIR data publication as well as data reprocessing and re-evaluation – possibly at the site of data creation. Our interest group (IG) “High-performance Computing in Earth System Sciences”, embedded within NFDI4Earth, will push these topics in the ESS sector.

This paper of our IG is intended as a concept paper to encourage both researchers and infrastructure providers to engage in joint discussions on the further design of work processes, future service offerings and their interaction on the background of the FAIR principles. The NFDI4Earth IG High-performance Computing in Earth System Sciences will further promote these discussions in NFDI4Earth and bring them into the framework of the entire NFDI.

2 General Situation, IG Activities and Scope of This Paper

The statements made above (Section 1) come in a time where, e.g., the analysis of large-scale remote sensing data becomes an everyday business due to the rising amount of freely available high-resolution satellite products (e.g. from Sentinel satellites). The strongly grown need for HPC in geoscience is reflected in the NFID4Earth proposal where HPC-related topics are mentioned throughout different task areas and measures. The exact concepts for data FAIRness in this sector, however, still need to be elaborated. There is general agreement that moving extremely large datasets must be avoided, but this raises further questions: How can data on large file systems or tape archives of supercomputers be published, and how can data users be possibly allowed to process the data at the facilities where they are stored? These questions are just starting to be addressed within NFDI4Earth and within NFDI in general, and a concept for nationally or internationally harmonised solutions needs to be developed. Likewise, efforts to make large-scale ESS data-processing workflows reproducible are in their infancy. These must be worked out, – based on FAIR data and reproducible processing steps, supported by adequate hard- and software.

Different HPC centres are starting to tackle these problems by establishing additional services for users that address single aspects of the FAIR principles for HPC data. One example is the harmonised development of metadata annotation services within the InHPC-DE project of the 3 GCS computing centres (https://www.gauss-center.de/news/pressreleases/article/project-inhpc-de/). Within other German HPC collaborations, similar efforts have been started, and for sure NHR and the Gauss Alliance will play a major role in future efforts. One basic aim is to store metadata with simulation data, to obtain DOIs for such data, and to make the data accessible on the original file system without moving them. Such services are complemented with domain-oriented data federations, which are able to offer much more specialised functionalities and services. A prime example for this is the Earth System Grid Federation (ESGF) service at DKRZ. This service not only brings together massive amounts of globally distributed standardised and quality assessed data, but also holds rich, standardised metadata that enables informed re-use. Data services are complemented by next-generation processing services with easier access (e.g. via Jupyter Notebooks at many computing centres). On time scales of the current NFDI4Earth funding period, one can expect to harmonise these efforts and thus offer capacity for publication. Also, reprocessing power for ESS applications can realistically be made more easily accessible. A seamless experience with a German, NFDI-based "data/metadata federation" and a federated "cloud-like" computing infrastructure will be much harder to achieve in particular in the HPC sector. Nonetheless, this idea shall remain an important vision. The longer-term and large-scale goals of the NFDI initiative are set – to a large degree – by the NFDI-Verein (NFDI e.V.). The NFDI-Verein has started work in four cross-cutting sections, allowing members of all funded consortia to contribute: "(Meta)daten, Terminologien und Provenienz", „Common Infrastructures", „Training & Education", and „Ethical & Legal Aspects". The work of these sections shows a thematic overlap with the work of our interest group, so that synergies will be created by active participation of IG members in the sections. Our IG will concentrate on two pillars:

1. "(Meta-)data Formats, Interoperability and Reproducibility" – i.e., harmonisation of annotation approaches and possibilities to annotate big datasets with metadata right at the HPC centres; selection of annotation schemas/procedures so as to allow for interoperability and reuse, and to enhance reproducibility. This topic is addressed in collaboration with the section "(Meta)daten, Terminologien und Provenienz".
2. "Federated Access, Findability and Accessibility" – i.e., inclusion of (meta-)data in federated data catalogues or systems for federated access, but also common or simplified access to on-site data analysis facilities to analyse data where they are currently stored. This pillar overlaps with actions of the section "Common Infrastructure", harmonising existing and future technical backbones of NFDI consortia towards a cross-disciplinary multi-cloud federation.

The aim of this concept paper is to highlight the current limitations of FAIR HPC data handling, and to provide ideas on how corresponding services and infrastructure could be improved to support the FAIRification of large geoscientific data sets. In order to illustrate current limitations, we will sketch two use cases from HPC "power users" in the following section. Use case 1 focuses on HPC simulations and data analysis for ice sheet modelling, while use case 2 addresses a high-resolution simulation workflow for Earth system models (atmosphere + ocean). Both use cases describe state-of-the-art workflows that use computing resources for geoscientific simulations and data analysis. Thereafter, we will analyse these two use cases with respect to shortcomings in the current infrastructure landscape, provide recommendations for future developments, and outline possible bottlenecks in putting such developments into practice.

3 Description of Use Cases

Use Case “Ice Sheet Simulations” (Angelika Humbert)

The aim of this use case is to sketch the future procedure of preparing simulation data of ice sheets for FAIR data provision. We anticipate three stages: (1) simulation run-time, (2) data post-processing and (3) data dissemination.

Simulation Run-time

In stage 1, there are two challenges to be considered: (i) data assimilation and (ii) on-the-fly analysis. For data assimilation, the key issue is how satellite remote sensing products will be provided for simulations. The major challenge here is that big input data needs to be acquired and handled. Just as an example, some 250 scenes are covering the margins of the Greenland Ice Sheet with Sentinel-1, with a repeat cycle of 6-days. With the ice sheet code being run on HPC platforms and data storage on cloud systems, efficient data transfer and storage is important. In this particular use case, data assimilation is done by a level set method, and synergies with the level-set methods for evolving the margins of ice sheets may be found. Within the assimilation process, bottlenecks in HPC-Cloud interaction must be avoided and hybrid parallelisation needs to be employed. On-the-fly data analysis – sometimes in interactive manner – has been another long-standing topic when doing production-quality predictions on HPC systems. In ours as in other fields (e.g. accurate and reproducible aeroplane behaviour prediction), simulations must possibly be checkpointed or even halted for semi-automatic or manual health checks. The on-the-fly analysis will also have to compute standard scalar quantities. On the long term, analyses shall further be extended to detect model shocks, which are typically happening when types of forcing are switched or when geometry evolution is switched on abruptly. A major requirement for any on-the-fly analysis is that it does not negatively affect load balancing or more general scalability.

Data Post-Processing

Stage 2 is starting with a basic quality check in case no on-the-fly analysis was conducted. This stage will consist of three blocks: AI-based analysis, level 2 (L2) product generation and computation of further metrics. AI-based analysis of simulation output is in its infancy, yet already yielding very promising results. It can be used to understand simulation properties (together with computed metrics), to compare the observations and to assess the reliability of the simulation.

The second part in stage 2 is concerned with L2 data product generation. This ranges from adding standard quantities (which are sometimes also computed during run-time) to adding quantities in drainage basins of ice sheets. Thus, the topic of attribution is addressed, and questions like ‘Why has the ice sheet lost mass? Due to changes in snow accumulation or melt? How much ice was lost by acceleration?’ can be answered. This is of particular interest as also satellite observations are now getting into the level of attribution, which allows for further assessment of simulation results. The basic idea behind using all this is expressed in the question: ‘Is the ice sheet in simulations losing the mass due to the right reason?’ For the glaciological community, further in-depth analysis is useful. To this end, different metrics are computed, but they are not yet standardized. The efficient computation of these metrics is the most pressing topic on the software and data science side. While metrics are developed by theoreticians, the computation of these measures needs to be done in parallel. On the data-management side, some of the metrics might actually be treated as metadata in a modern FAIR Research Data Management approach.

Data Dissemination

While Stages 1 and 2 depend crucially on access to computing systems, Stage 3 is concerned with providing the data to stakeholders and scientists. The demands of both groups differ widely. Given that ice sheet simulation results are rarely used as input data for other ESS disciplines, the data provision in this field may differ substantially from atmospheric or oceanic data sets. Depending on user demands, individual data portals would have to be built, providing geo-location- and time-selection-based downloads. Filtering immense data sets is also a common requirement in order to access the fields of interest only. As an illustrative example from our work, one should be able to click on a geographical location, e.g. a glacier, and be able to get a dashboard including its major attributes as velocity over the simulation time, retreat/advance, discharge, contribution to global sea level change, and so forth. Typical further stakeholder requests include on-line analysis of effects (e.g. of ice sheets from different regions on the sea level) or an alert if something happens in an area of interest. At this time, it is a completely open question how data residing on supercomputer file systems can be broadly provisioned fulfilling such requirements and the FAIR principles, and how persistent identifiers can be assigned to them with the possibility of also addressing sub-datasets.

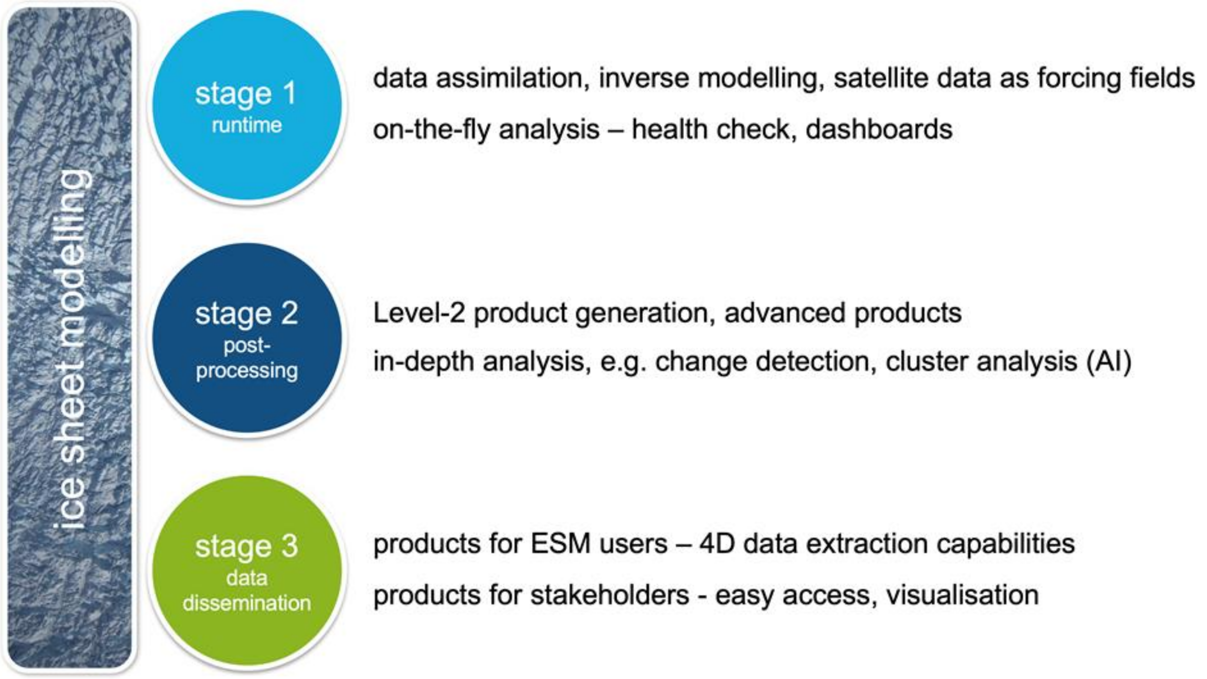
ice sheet modelling
stage 1
runtime
data assimilation, inverse modelling, satellite data as forcing fields
on-the-fly analysis – health check, dashboards
stage 2
post-processing
Level-2 product generation, advanced products
in-depth analysis, e.g. change detection, cluster analysis (AI)
stage 3
data dissemination
products for ESM users – 4D data extraction capabilities
products for stakeholders - easy access, visualisation

Figure 1: Sketch of the workflow from ice sheet simulations to data provision.

Use Case “High Resolution Simulations of the Earth Climate System” (Nikolay Koldunov)

This use case is described in the same manner as the previous one, focusing on simulation run-time, data processing and data dissemination. It involves a typical workflow of post-processing the high-resolution simulations performed with Earth/Climate System Models (ESMs) or their individual components (e.g. atmosphere- or ocean-centric components).

Simulation Run-time

An ESM requires periodical input of information about boundary conditions. In the case of coupled ESM simulations this can involve relatively small ASCII files, containing concentration of CO_2, aerosols and similar parameters. In the case of simulations focussing on atmosphere or ocean, the amount of information is much larger, since the model requires updates about the state of the boundaries several times per day. Usually the boundary conditions are taken from reanalysis, and their data volume depends on the spatial and temporal resolution of the reanalysis product. The largest size of the fields which are necessary to run the FESOM2 ocean model with the highest resolution global reanalysis product is 270 GB per year and is used in practice today (ERA5). These data have to be downloaded, possibly pre-processed and managed.

During the model simulation one has to periodically check the model state to make sure the simulated fields are within realistic bounds and not, for example, drifting away too much. This can be done either “by hand” (scientist periodically performs monitoring) or automatically (periodical execution of monitoring scripts). These tasks usually do not require very heavy data analysis capabilities, as the checks could be rather simple, and done on reduced (e.g., interpolated) data. This approach is named “off-line monitoring”, as the analysis is performed on data serialised on disk. Currently the concept of “on-line monitoring” becomes more popular, where data analysis is performed alongside the model run, possibly on data that are still in the memory. The applications of such an “on-line monitoring” are triggering state aware model output (e.g. more frequent output when a storm of specific characteristics is present over some area), on-line data analysis on the fields with higher temporal frequency (e.g. aggregating values every time step), applying ML techniques for model parametrisations, ensemble pruning based on some criteria, and so on. However, the majority of model simulations still uses classical “off-line monitoring”.

Data Processing

ESM simulations are quite demanding in terms of data storage. The volume of the data output depends on model spatial resolution, temporal frequency of the output and the number of output variables. In realistic applications, it currently ranges from a few MB per year to a few TB per year, but those volumes will be constantly growing, as the spatial resolution of model components continues to increase. The most popular format for storing ESM data is netCDF, while GRIB is also sometimes used for atmospheric modelling and numerical weather prediction.

After the model simulation is finished, some initial post processing tasks might be necessary. Those include computation of standard derived variables, different types of interpolation and conversion of data to different formats, or adding some metadata. The tools that are used include bash scripts, python or other scripting languages with scientific modules, climate data operators (cdo) and netCDF operators (nco). The critical part for which HPC centres are less prepared, however, follows after these standard tasks: Large-volume simulation data should be available on adequate storage systems over long periods of time (months to years) before they can be put to the tape archives. Within this period, researchers want to evaluate the data with exploratory, ideally interactive analysis. To be “FAIR”, this should not only be possible for users of the particular HPC system where the data are stored, but also for external scientists interested. This is crucial not only to counter the “reproducibility crisis” but to make proper scientific progress: scientists should be able to develop and apply new diagnostics, try different options, select different time scales and regions, and quickly visualize the results. This also should happen in the regime that is close to interactive, as constant breaks in the exploratory workflow limit the possibility of trying many scientific ideas and hinder creativity.

To complicate this, big data post-processing is often not realistically possible on one serial-processing node any more. Demands have strongly risen, and one has to use parallel algorithms and work on one or several HPC nodes to process today’s large-volume output. For the FESOM2 ocean model, this involves using Jupyter notebooks together with the python xarray library with dask support, enabling parallel processing.

All in all, we see the following challenges in our post-processing and data analysis approach:

- Access to HPC resources. Exploratory data analysis means that the user needs time to evaluate results, come up with new ideas and write new code. During that time, the computing resources are not used. The batch scheduling systems of HPC centres currently are not flexible enough (or not tuned) to accommodate this work pattern and allow for long, interactive and partially idle multi-node jobs. They rather fill computing-cluster nodes with highly-efficient automated jobs, as they appear in simulations, but not in the necessary post-processing efforts.
- Fast disk and data formats. Most of the time the bottleneck in ESM data analysis are the input/output (I/O) operations. This becomes especially important for parallel data

processing, when all the data cannot just be put to the memory at once, but has to be read (less frequently written) from (to) disc in relatively small chunks. Efficiency problems in parallel I/O and other limitations may even lead to the consideration of novel file formats such as Zarr within our community, which however brings its own challenges in addition.

Data Dissemination and Sharing

The amount of resources used to compute high resolution ESM simulations is enormous, and it is very desirable that results from those simulations are analysed and reused as much as possible. The most effective way to do this is to share the data with other groups in the "FAIR" spirit, and provide tools and infrastructure to perform scientific analysis. Currently, data sharing is done in an archaic way, and most of the time comes down to either providing a path on the HPC system where the data are stored (if the potential user of the data has an account on the same machine), copying to a different HPC centre, or sharing files on-line through ftp. In some cases, simulations are shared via more advanced protocols like OPeNDAP, gridftp/Globus, uftp or cloud storage. However, this is usually done for already well-established datasets, as it may require a lot of preparation and interaction with the respective HPC centre, for example to adjust the access rights.

In order to comply with FAIR principles in case of ESM data on HPC systems, re-usability (corresponding e.g. to accurate descriptions and good file formats) needs to be complemented by a sufficient support for the "F, A and I" principles:

- Findability: The users should be able to expose the data themselves, without involving complicated procedures from HPC centres. The data should sit not in some obscure on-line catalogue behind the centre's VPN, but be discoverable through google search. Part of findability is to be able to provide an initial set of precomputed standard visualizations for the data, so that potential users can quickly evaluate the dataset.
- Accessibility: User/web applications should be able to visualise/download data/portions of data "programmatically", i.e. via standard APIs such as S3. Barriers for access (logins, ...) have to be minimised. Storage formats have to be standardised, e.g. based on cloud/object storage approaches and appropriate file formats such as Zarr.
- Interoperability: Being able to expose your data, so that other users can download portions of them through API would be a great step already. The next level is to do data analysis on those data for internal (on the same HPC) and external users.

To put this together in a concise manner, what we need is something that looks and acts as a storage "mounted" to our computer, so that we can either copy data from it, or do some operations on the data.

4 Challenges from Gap Analysis

From the use cases discussed above we identified two types of challenges for NFDI4Earth and its HPC-centric activities: i) Challenges arising during the immediate usage of computing systems (e.g. HPC simulations, post-processing) and ii) "FAIR" challenges appearing when sharing and re-using (big) data created at (super-)computing centres. These are laid out below as a summary.

Challenges in Immediate HPC/Computing-System Usage

We identified challenges in the HPC usage and related components in almost every step of typical ESS use cases. Below, we list them by category.

Challenges in (big) data management:

- efficient handling of large (as of number of files and size) and dynamic (several updates per day) data collections
- efficient methods for gathering and selection of data from large data sets from different sources for analysis or processing

- fast and scalable access to data on different resources (e.g. data in the Cloud, analysis on HPC)
- long-time provision of data (month to years) for analysis (i.e. not on tape)

Challenges in application monitoring and steering during run-time:

- establishing simulation monitoring for analysis and quality control
- move from regular automated or manual checks to sophisticated on-the-fly analysis
- ML capabilities for on-the-fly analysis needed (e.g. for model parametrisations)
- monitoring frequency (might depend on current application state, e.g. storm in simulation over some specific region)
- user interaction with the application for steering

Challenges in system and data access:

- have as little barriers in terms of logins/accounts as possible
- fast and scalable (parallel) data access for interactive data analysis
- new community standards like Zarr – small and numerous chunks on HPC file system

Challenges in data processing/analysis:

- interactive data exploration or analysis on single or several nodes needed, with increasing parallelisation on the long term
- on-line quality control needed for sustainability, reusability and reproducibility in some cases

Challenges for Data FAIRness, Data Sharing and Re-use

Earth system sciences live from data sharing. Therefore, some of the challenges are connected to the provision of (FAIR) data that were created on HPC and are too large to be moved away from these systems.

Challenges related to metadata and findability:

- meaningful metadata needed; may contain some “hints” (e.g. global characteristic quantities of simulations, or geo-locations) for data analysis
- direct publication of large data sets (which still reside e.g. on HPC storage) by the user needed
- interface to find data must be public, not “behind a VPN”

Challenges related to data access:

- provide data for different user groups (scientists, stakeholders) with different demands
- allow for selection/download of sub-datasets based on time, geo-location, field, etc.
- different levels of information (from overview to detail information, e.g. when a region is selected)
- direct data access from the outside with API (e.g. S3) or a file-system-like mechanism

Challenges related to interoperability:

- standardised and meaningful data and metadata formats

Challenges related to in-place analysis for large datasets (reusability):

- interactive analysis capabilities necessary (e.g. Jupyter) for re-users which might not be users of the HPC system where data are stored

Other challenges:

- capability to subscribe to data sets (“alerting”) for information on updates or data events

5 Recommendations for Future Development (General / Specific to Pillars and Challenges Identified in Section 4) and Outlook

With this section, we close our paper, recommending some initial development directions for NFDI4Earth, which partially should apply to NFDIs in general. First, we give a few general directions; afterwards, we give recommendations specifically derived from the findings in Section 4 (separated by the two pillars of our work). We end the paper with an outlook on how the IG will monitor progress towards its goals.

General Recommendations

Our idea of future data and compute services is based on FAIR data, but also computing approaches which consider the "FAIR spirit". Although it is not trivial to define what this means in the HPC/computing sector, it certainly includes on the compute part:

1. An infrastructure registry for NFDI4Earth to find matching computing resources;
2. Approaches to access computing resources through a "Single Sign On" (SSO) via central authorisation and authentication infrastructures (AAIs) provided on national or European scale, in particular the AAI(s) that the NFDI e.V. is considering;
3. Approaches to provide (and also – on large scale – easily account for and bill/pay for) this computing access with a border-/seamless experience to the user, which boils down to making distributed computing comply with local terms of usage;
4. Approaches to interconnect computing systems (interoperability); and finally
5. Approaches to provide robust computing environments (e.g., via containerisation) to warrant re-usability of data and reproducibility of processes.

Recommendations specific to Pillar "(Meta-)data Formats, Interoperability and Reproducibility"

Use-case oriented recommendations on actions within this pillar, derived from the needs brought out in Section 4 (above), are given below:

6. Establishment of a limited number of proper data and metadata storage schemes and recommendations for annotation in collaboration with the Research Data Alliance (RDA) and other standardisation bodies; this shall foster interoperability and reusability for fellow researchers, and a reproducibility of (meta-)data-driven workflows. In particular this shall manifest in:
    a. NFDI4Earth recommendations of ≤ 10 general storage formats in ESS;
    b. NFDI4Earth recommendation of ≤ 10 metadata formats in ESS;
    c. NFDI4Earth recommendations of usage of these formats throughout the data lifecycle;
    d. Support for these (meta-)data formats at all computing centres within NFDI4Earth.
7. Establishment of standardised software environments – considering containerisation – at computing centres, and of a central national documentation within NFDI e.V. or NFDI4Earth
8. Issuing Recommendations on Research Software Engineering and Software Packaging with focus on Reproducibility
9. Issuing Recommendations on Quality Control
10. Political work towards viable solutions across state and national borders

Recommendations specific to Pillar "Federated Access, Findability and Accessibility"

Finally, we have some recommendations derived from Section 4 which pertain to the pillar "Federated Access, Findability and Accessibility":

11. Issuing NFDI4Earth recommendations on efficient data access and transfer for Big Data (mechanisms, interfaces, technical requirements)
12. Removal of administrative barriers regarding federated access to data/computing resources
    a. Establishment of AAI/Single Sign On solutions to data and compute infrastructure following actions of NFDI e.V.
    b. Establishment of portals and aggregator portals giving free, unrestricted access to Open Data, without the necessity of moving data from their original storage

    c. 1...2-yearly user surveys to find users which have no compute/data service supply and remove the respective gaps

13. Establishment of data-portal and storage solutions allowing for geo-location-, time- and layer-/field-based selection and download of data
14. Establishment of standardised metadata and data interfaces to institute- or subdiscipline-centric infrastructure, and between HPC and other computing infrastructure
    a. Further development of standardised metadata interfaces, starting with OAI-PMH
    b. Advancing usage of e.g. S3 API as an important data-access interface
    c. Efficient data-transfer mechanisms
15. Clarification and NFDI4Earth index of long-term (in particular HDD-/SSD-based) storage solutions
16. Work towards novel usage models at computing centres, allowing for interactive computation and checkpointing/interactive HPC schemes in an efficient way
17. Concepts for enhanced services reacting on data events (e.g. alerting, instrument operation)

Clearly, this body of recommendations cannot be fully implemented within the first 5 years of NFDI activities. However, a group-internal review every 12-24 months will show current status and changes within the previous 18 months so as to monitor progress on the targets stated above. On the same timescale, updates of this concept paper shall be issued (and published on Zenodo as new versions of the original document). With this consideration, we close our concept paper and repeat the invitation to participate in our IG to everyone in NFDI4Earth.