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/

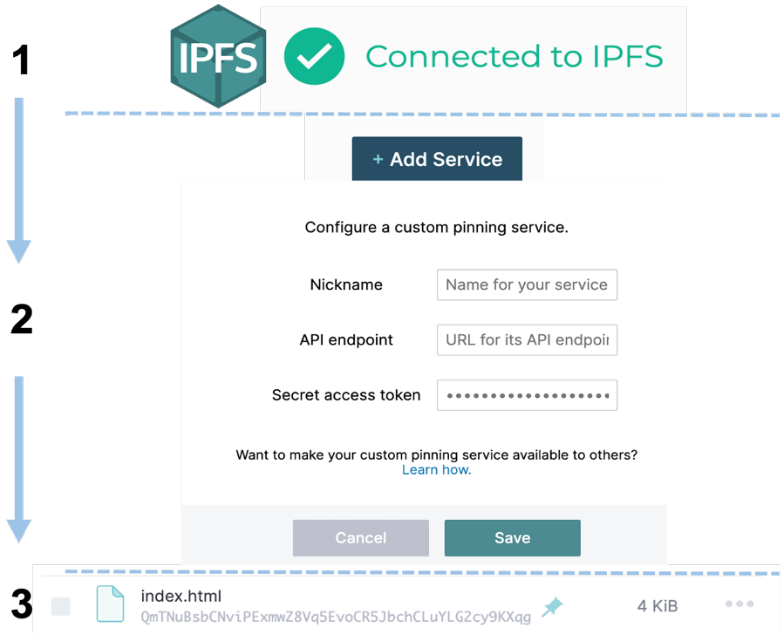
[2] https://eurec4a.eu/

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.

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