Investigating Decentralized Alternatives to Collaborative Long-term Research Data Preservation Infrastructure

Belouin, Pascal

pbelouin@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Pham, Kim

kpham@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Hennicke, Steffen

shennicke@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

The sustainability, preservation, and long-term availability of research data is a growing concern for academic institutions. Several centralized research data storage and hosting solutions are currently available globally to mitigate this issue, where research data archiving and access is guaranteed by a number of centralized entities, often hosting research data for several academic institutions at once.

In such a centralized consortial hosting model, each institution provides access to a particular collection of datasets. Coordination and cross-search capabilities between these centralized repositories is ensured by initiatives such as EUDAT (EUDAT, 2022), through a number of ways such as OAI-PMH (Open Archive Initiative Protocol for Metadata Processing, 2022), CSW (Catalog Service for the Web, 2022), or ad-hoc REST APIs. This approach relies on the existence of the aforementioned institutional entities, and on their ability (and willingness) to maintain the infrastructure that provides access to these data in the long term.

Over the recent years, a move for the creation of decentralized infrastructure has been observed, with initiatives such as IPFS, or more generally projects categorized under the all-encompassing "blockchain" umbrella. A notable initiative at the intersection of blockchain and academia is for instance the bloxberg consortium, spearheaded by the Max Planck Digital Library, and which aims to *"advance science with its own blockchain infrastructure and to enable society as a whole to secure data with the reputational proof of research organizations worldwide" (bloxberg Consortium, 2022)*.

In this context, we have been investigating a number of ways in which such a decentralized model could be implemented. A number of technical possibilities have been explored, such as the development of a proof of concept for a research data repository based on SPARQL-queryable linked data, stored in JSON-LD using IPFS technology. More recently, a new solution that seems to offer most of what we think would be the requirements for such an infrastructure advantages is the Decentralised Knowledge Graph developed by Trace Labs, which they describe as an *"immutable, queryable and searchable graphs that can be used across Web3 applications. You can additionally apply standardized technologies such as GS1 EPCIS, RDF/SPARQL, JSON-LD and other W3C and GS1 standards out of the box" (OriginTrail, 2022).*

During this short talk, we therefore aim to compare existing long-term research data preservation solutions to potential fully decentralized alternatives, explore the pros and cons of the latter, and present the current state of our prototype for such an infrastructure.

This investigation was motivated by our belief that some of these novel approaches to data storage infrastructure could present a number of advantages to some of the problems entailed by the long-term data preservation solutions that currently exist.

The prototype is a decentralized and distributed architecture that aims to do away with the reliance on a small number of centralized entities for storing and accessing the data. In the context of a decentralized & distributed storage model, all datasets are distributed across the network in such a way that the loss of one or more nodes does not result in the loss of data. This brings about several benefits such as censorship resistance, immutability of the data, the removal of a single point of failure, and the removal of a "direct" need to maintain the actual infrastructure necessary for long-term storage and availability. Instead, the responsibility of maintaining each node is delegated to node operators, dependent on the incentivisation model underlying such an infrastructure, which can for instance take the form of direct remuneration of node operators based on uptime or volume of data hosted, or through participation in a consortium of institutions agreeing to be part of the network of nodes.

In addition, such a solution also has advantages, such as facilitating the interoperability of the research datasets stored in such a way, as although storage is distributed, such an infrastructure has properties similar to a centralized data storage solution: In other words, instead of a collection of discrete repositories, research data could be stored in what can ultimately be considered, from the outside, as one cohesive data repository. However, It also comes with a number of new challenges inherent to the technological stack underlying it, regarding, for example, issues of identity management, data ownership and access, and the incentivisation of data hosting and node maintenance. Furthermore, the institutional ramifications of adopting such a model also need to be addressed: Institutions participating in such an initiative could do so either by providing funds to incentivize storage, or by hosting a node.

Bibliography

EUDAT - Research Data Services, Expertise & Technology Solutions . (n.d.). Retrieved 4 November 2022, from https://www.eudat.eu/

Open Archives Initiative Protocol for Metadata Harvesting . (n.d.). Retrieved 4 November 2022, from https://www.openarchives.org/pmh/

Catalog Service for the Web (CSW)—GeoNetwork opensource v3.10 GeoNetwork Documentation . (n.d.). Retrieved 4 November 2022, from https://geonetwork-opensource.org/manuals/3.10.x/en/api/csw.html

Mission | bloxberg . (n.d.). Retrieved 4 November 2022, from https://bloxberg.org/discover/mission/

OriginTrail—Decentralized Knowledge Graph (DKG) . (n.d.). Retrieved 4 November 2022, from https://docs.origintrail.io/general/dkgintro