WORKSHOP

Decentralized Storage for Digital Preservation

Marco Beck, Cornelius Ihle, Moritz Schubotz, Patrick Sahle, Bela Gipp,
Stefan Strathmann & Markus Schmalzl

Agenda

1. Introduction
   (Marco Beck)

2. Decentralized Storage / InterPlanetary File System (IPFS)
   (Cornelius Ihle)

3. ~~NFDI working group on common infrastructures (Dr. Markus Schmalzl)~~

cancelled, due to short-term prevention

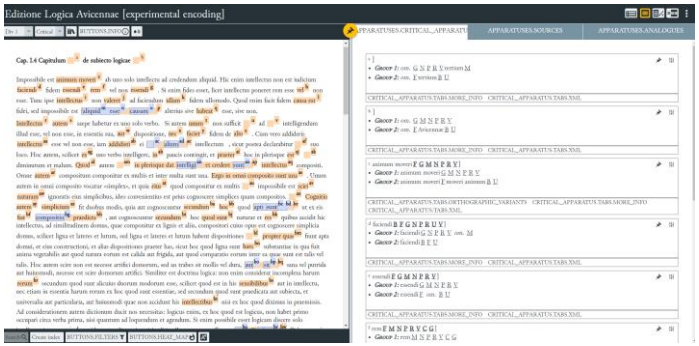
4. Round table (Discussion/All)

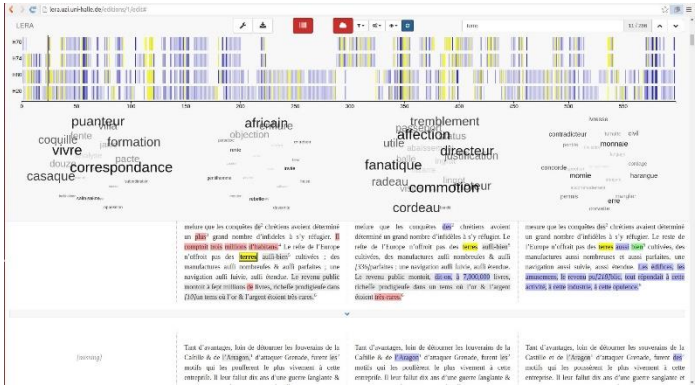
1. Introduction

Marco Beck

Digitization as a basis for further research

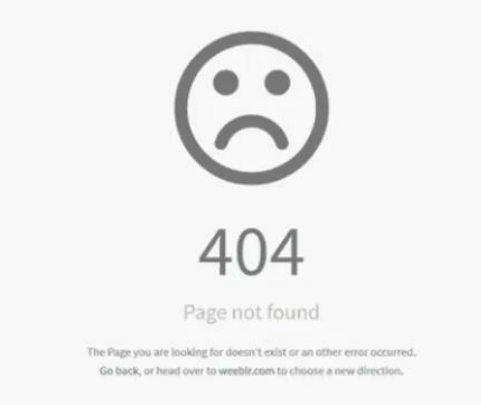
- Paradigm shift through digitization in humanities research as well as in the art of editing.
- Digitized objects and editions from libraries, archives and museums are the basis for any investigation and further research.
- The goal: to make **digitally edited sources** available **online** in order to make them **searchable** and usable for **further research questions**.
- For example, **interactive links** with **external archive material** or **analyze** and **visualize** the **text phenomena**.





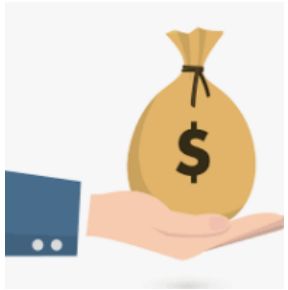
Fates and sustainability issues

- Digitization of the editions is mostly done on a **project basis** and is **funded** by **third parties** -> where to after the **end of the project**
- **Removal**
- **Technical** und **Organizational** break
- **Transformation**, downgrading, integration
- Many older digital editions are **no longer accessible**
  - Thomas Raddall Electronic Archive Project
  - Streuvels Edition
  - Alcalá Account Book Project



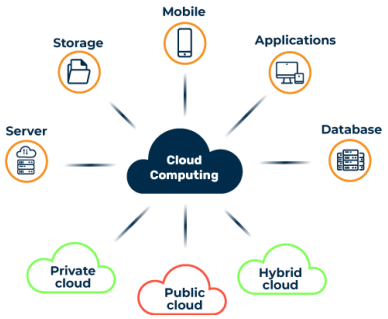
Challenges / problem areas

- **Problem** area: **sustainable storage** and **long-term operation**
- So far, **no reliable financing** models for the long-term operation of digital editions or edition platforms.
- Continued use of data, e.g., through **generic APIs** and variable layers of annotative information, e.g., in an **IIIF scenario**, poses a major challenge



Self-Hosting vs. Cloud Computing

- **Cloud computing**, which improves **availability** and **redundancy**, while traditional self-hosting by individual institutions could become less important as institutions close, restructure or also lose the financial resources for long-term operation.
- **Cloud computing** as a driver of **flexibility** in **pricing** and **scalability**.
- Disadvantage that there is a **single point of failure**, the **hosting company** itself.
- Both **cloud providers** and **institutions** potentially subject to **government censorship**.



Possible further approach

Sustainable storage and the long-term operation of digital data and services would be **decentralized storage** within a **content-addressable peer-to-peer method**.

2. Decentralized Storage / InterPlanetary File System (IPFS)

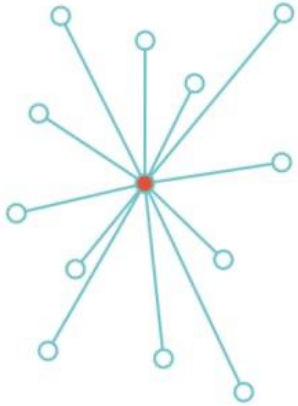
Cornelius Ihle

Overview

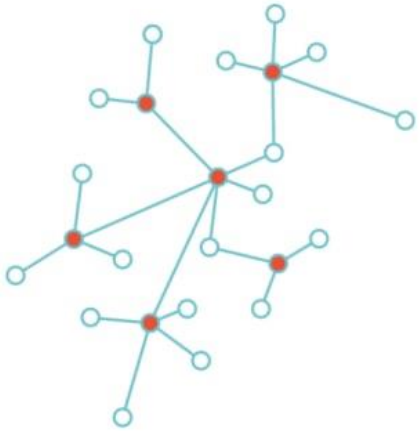
- Decentralization
- Content Addressable Storage and CIDs
- Experiments and Vision
- DHT
- TEI and Digital Editions
- Incentives

Decentralization

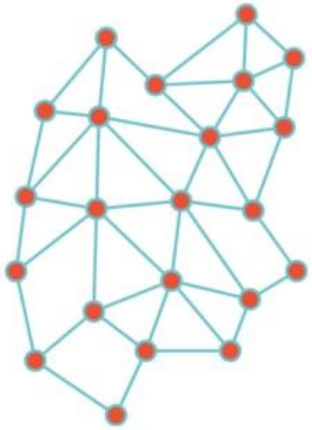
Building Resilient, Secure, Transparent, and Trustless Systems



Central

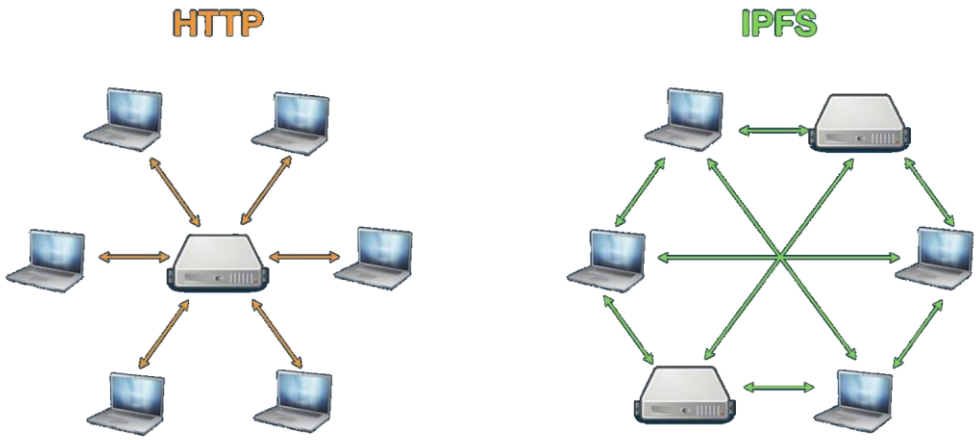


Distributed



Pure Peer-o-Peer

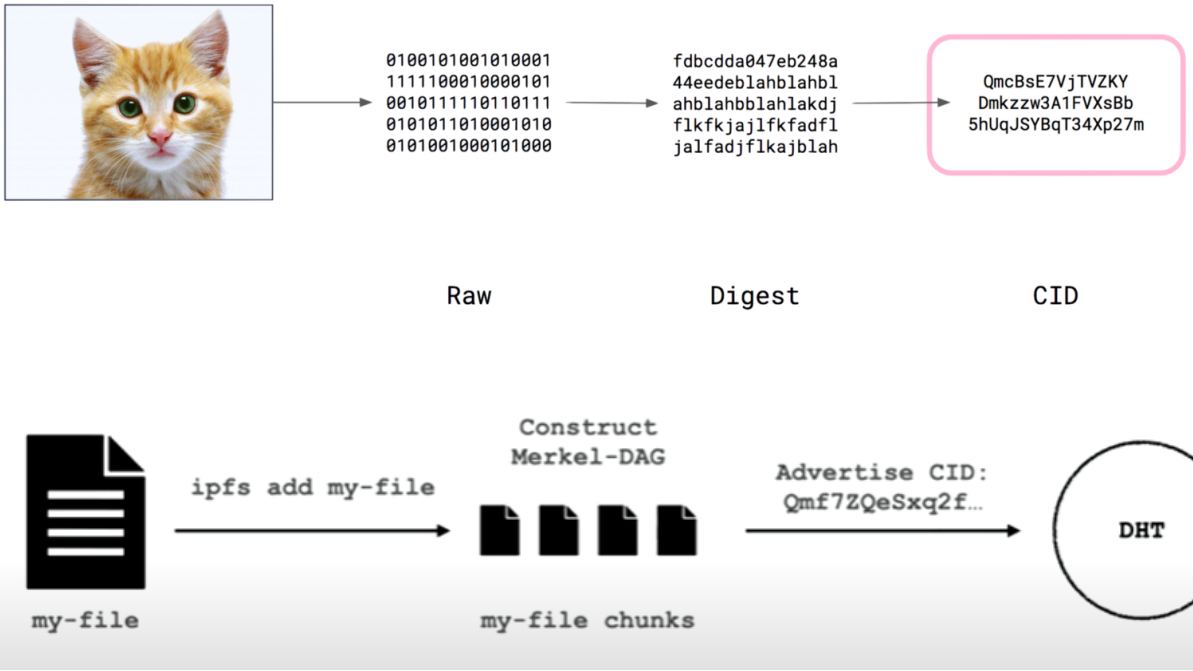
An Alternative to Location Dependend Storage



Immutable, Verifiable, Trustless, Permanent*Identifiers

404 Page Not Found

CID



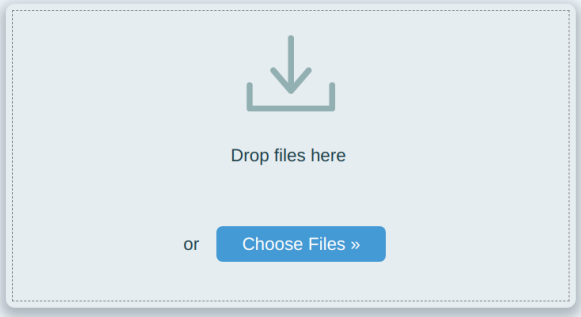
<multicodec><multihash-algorithm><multihash-length><multihash-hash>.<protocol>.<gateway>

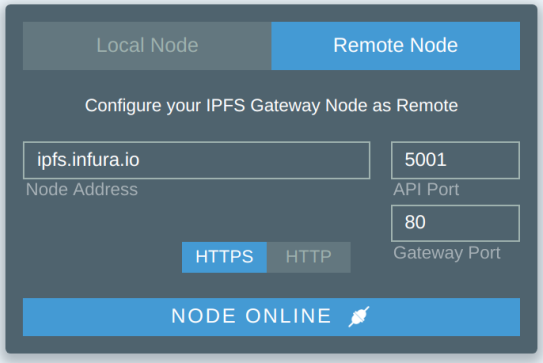
https://bafybeiceu7x3e2hacjrkqpsmvzijag6jfpcbkdcozdhxh5zzljorwylrxe.ipfs.infura-ipfs.io

DHT – Distributed Hash Table

- Sharing CIDs with others
- A distributed key/value store
- Keys can be PeerIDs or CIDs
- PubSub would be an alternative to a DHT to announce new content in the network

Using IPFS Today





0GitHub-Mark.png
(image/png) - 7110 bytes, last modified: 6/6/2019

eu.jpg
(image/jpeg) - 308945 bytes, last modified: 6/20/2019

index.html
(text/html) - 789871 bytes, last modified: 6/17/2019

ipfs.txt
(text/plain) - 20 bytes, last modified: 6/7/2019

mensage_nano.py
(text/x-python) - 1107 bytes, last modified: 6/14/2019

script
(n/a) - 71 bytes, last modified: 7/12/2019

update.txt
(text/plain) - 30 bytes, last modified: 6/9/2019


```
[
  {
    "path": "ipfs.txt",
    "hash": "Qmdjv6rFs5YdmvHDvwc4b1yAMBAuifdrWs51mq2UTUd9kn",
    "size": 28
  },
  {
    "path": "0GitHub-Mark.png",
    "hash": "Qma1kwFL54avBTAgzUzUHFjFu71VLh4AR9ytDczMe1S4Ef",
    "size": 7121
  },
  {
    "path": "script",
    "hash": "QmcjN61mUL944v5tmPQhAYua31X6kPavV73K2SctnYCmdu",
    "size": 79
  },
  {
    "path": "update.txt",
    "hash": "QmRfzsoxDWv9zMF2kuvAL4kgtu8wkST7xgrEMrfWYV4rTg",
    "size": 38
  },
```


https://anarkrypto.github.io/upload-files-to-ipfs-from-browser-panel/public/

https://web3.storage

Storage: *2.8 MiB* of *1 TiB* used

Search my files ›

1 TiB 1 TiB+

Stored files 2.8 MiB

Need more free storage? Submit a request

Upload more files

Files can also be uploaded directly using this web interface, as an alternative to using the API

Upload Files

Files

Upload +

Search for a file

Refresh

Sort By

Name	CID	Status	Storage	Size	Date
Upload at 2022-05-25T09:55:21.097Z	bafkr...hvngsply	Pinned	Stored (6)	217.01 KB	25.5.2022
Upload at 2022-05-25T09:28:42.684Z	bafkr...j3zjnfge	Pinned	Stored (3)	4.47 KB	25.5.2022
aulendorfer.xml	bafyb...h55xpjxa	Pinned	Stored (6)	1.49 MB	11.3.2022
beth.xml	bafyb...4jickaa4	Pinned	Stored (6)	658.56 KB	11.3.2022
eu.xml	bafyb...oq6xy7dm	Pinned	Stored (6)	446.4 KB	11.3.2022
diary.xml	bafyb...p7qqoncu	Pinned	Stored (6)	7.15 KB	11.3.2022
derSturm.xml	bafyb...zwlzlhpe	Pinned	Stored (6)	9.1 KB	11.3.2022

Archived Data on IPFS

- https://en.wikipedia-on-ipfs.org/wiki/
- https://ipfs.io/ipfs/QmSnuWmxptJZdLJpKRarxBMS2Ju2oANVrgbr2xWbie9b2D (Apollo)
- https://awesome.ipfs.io/datasets/

An Alternative to Location Dependent Storage

- Automatically route to **alternative** poviders (DHT)
- Automatically connect to **alternative** addresses (multiaddresses)
- Automatically verify content by its identifier (CID)

- Higher latency for provide and retrieve operations

Accessibility and Availability

Long Time Archiving

- Funding
- Selfhosting/Participating

Technological Best Practices

- Static Content
- Client Side Rendering
- Single Page Applications

Vision

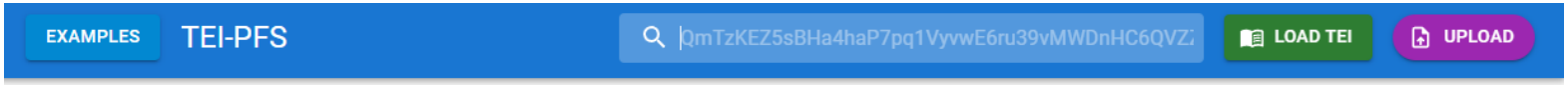
- Advantages

  Low Costs
  Low operational responisibilities
  Low technical know how (Plug and Play)

- Challenges

  Mirrowring existing databases in Web3
  Defining a best practice for the creation for digital editions

Experiments



Haag 28.4.20

Lieber Herwarth.

Dank für dein Brief, schön das ich nicht dem Reichskommissär zu schreiben brauche. Heute das Schreiben mit Gesellsch: Sturmfreunde prospekt.(Die „Gesellschaft der Sturmfreunde“ ging aus der Vereinigung des „Verein für Kunst“, des „Verein Sturmbühne“ und des STURM-Klubs hervor und übernahm vor allem die Planung und Ausrichtung der STURM-Bälle. Vgl. STURM-Bühne: Jahrbuch des Theaters der Expressionisten, Nr. 6, Mai 1919, S. 4. Ferner Pirsich, Volker: Der Sturm. Eine Monographie. Herzberg 1985, S. 340f.) erhalten. Schade das ich nicht in Berlin sein kann im Mai, so gern hatte ich es gesehen. Ich möchte lieber kein Mitglied der Gesellsch Sturmfreunde mehr sein, da ich doch so wenig in Berlin bin finde ich es zu teuer dafür 60. Mark zu zahlen. Erste Nummer von Der Sturm (Walden, Herwarth [Hrsg.]: Der Sturm. Monatsschrift für Kultur und die Künste 1 (1920).) erhalten gegen f(Die Abkürzung „f“ steht für die ehemalige holländische Währungseinheit Gulden, deren Name sich aus der Florentiner Goldwährung ableitet.) 1. remboursement für den Jahrgang. Da ich das nun so gezahlt habe braucht du es nicht auf mein Konto zu schreiben.

Ich bin sehr beschäftigt mit eine sehr grosse Bestellung für 3 enorme Glasfenster und ein Türpanele für eine grosse ganze neue Villa in der Nähe vom Haag. Meine Entwürfe sind direkt angenommen und in zwei Monaten wird alles fertig sein. Ich habe die ganze Leitung für das Glasbrennen etc:

Mit herzlichste Grüssen

deine JacobavanHeemskerck .

Ende Mai oder Anfang Juni wird ein Freund von mir ein junger artz für einige Tagen nach Berlin kommen und dann weiter nach Leipzig gehen. Er fühlt sehr für die ganze Sturm Bewegung. Bitte zeige ihm deine Sammlung, ich habe ihm ein Brief von mir an dich mit gegeben als Introduktion. Seine Name ist W. Zeylmans van Emmichoven

Three Witches, the Weïrd Sisters

https://bafybeiceu7x3e2hacjrkqpsmvzijag6jfpcbkdcozdhxh5zzljorwylrxe.ipfs.infura-ipfs.io

TEI

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="Q.01.19200428.JVH.01">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Jacoba van Heemskerck an Herwarth Walden, 28. April 1920, Den Haag</title>
        <editor>
          <persName ref="https://orcid.org/0000-0003-1208-6131">
            <forename>Anne</forename>
            <surname>Lorenz</surname>
          </persName>
        </editor>
        <respStmt>
          <persName ref="https://orcid.org/0000-0002-0953-2818">
            <forename>Torsten</forename>
            <surname>Schrade</surname>
          </persName>
          <resp>
            <note>Herausgeber</note>
          </resp>
        </respStmt>
        <respStmt>
          <persName ref="https://orcid.org/0000-0003-2423-7144">
            <forename>Marjam</forename>
            <surname>Trautmann</surname>
          </persName>
          <resp>
            <note>Herausgeberin</note>
          </resp>
        </respStmt>
      </titleStmt>
      <publicationStmt>
        <publisher>
          <ref target="http://www.adwmainz.de">Akademie der Wissenschaften und der Literatur | Mainz</ref>
        </publisher>
        <pubPlace ref="http://sws.geonames.org/2874225">Mainz</pubPlace>
        <availability>
          <licence target="https://creativecommons.org/licenses/by/4.0/">This file is licensed under the terms of the Creative-Commons-License CC BY 4.0 </licence>
        </availability>
        <idno type="file">Q.01.19200428.JVH.01.xml</idno>
        <idno type="uri">https://sturm-edition.de/id/Q.01.19200428.JVH.01</idno>
      </publicationStmt>
```


https://bafybeihybomlqn3to7n7bisul5pt6xl6encatxe7yybqbja5m6zwlzlhpe.ipfs.dweb.link/derSturm.xml

Decentralized Data Sharing vs Storage

Decentralized Data Sharing vs Storage

File Sharing (2001): Napster, BitTorent, BearShare, DC++ ..

Political risk of platform providers.

"Haftungsadressat des TMG ist der so genannte Diensteanbieter (Provider). Darunter fallen gemäß § 2 Nr. 1 TMG alle natürlichen oder juristischen Personen, die eigene oder fremde Telemedien zur Nutzung bereithalten oder den Zugang zur Nutzung vermitteln."

Incentive Mechanisms

Cooperation

Participation

Reputation

Service

Monetary

3. ~~NFDI working group on common infrastructures~~

cancelled, due to short-term prevention

Dr. Markus Schmalzl

Working Group NFDI “Common Infrastructures”

- The NFDI consortia have a variety of different and distributed information infrastructures, many of which are currently only loosely or insufficiently interconnected.
- The goal is to create a basic infrastructure based on the principle of the Research Data Commons (RDC). The RDC concept includes, for example, shared cloud services, an application layer with access to self-distributed high-performance computing, collaborative workspaces and terminology services, and a common authentication and authorization infrastructure (AAI).

4. Round table

Moderation Prof. Dr. Patrick Sahle, Marco Beck

Summary from the round table (1/2)

- Diskussion, ob voraussichtlich IPFS überhaupt noch in 50 Jahren existiert, respektive ist diese Technologie und der Ansatz langfristig vorgesehen -> Empfehlung ist zunächst einen parallelen Weg zu gehen.
- Man könnte heute bereits überlegen, inwiefern ein Link-Konzept aussehen könnte, wenn sich der Hashing-Algorithmus ändert. Eventuell könnte IPFS mit dem DOI-Konzept verknüpft werden.
- Bei LZA ändern sich auch die Objekte nicht nur die Hashing-Werte -> heute bereits sicherstellen, dass auch die Objekte mit der digitalen Edition verfügbar gemacht werden; keine externe Abhängigkeiten -> jedoch keine 100%-Sicherheit gegeben, dass auch die digitale Edition technisch aufrufbar ist (Beispiel Flash-Player).
- Bei Überarbeiten von wissenschaftlichen Editionen, dann würde die neue Edition mit einem neuen Hash-Wert gespeichert werden-> Lösungsansatz wäre das private/public-key konstrukt (Content-Betreibermodell); allerdings wer hat das Recht die Edition zu verändern bzw. eine neue zu erzeugen (organisatorisches Problem).
- Frage, ob es auch möglich wäre Daten per Schnittstelle aus IPFS zu extrahieren bzw. wiederzuverwenden. Ansatz wäre, IPFS search, welche als Suchfunktion anzusehen ist, mit welcher der Content durchsucht werden könnte, allerdings muss hierfür der Einstiegspunkt von IPFS bekannt sein. Ein globaler Verzeichnisdienst sowie das reine Nutzen von Daten ist heute noch nicht verfügbar.
- Eine komplette Applikation (Web-App) kann noch nicht in IPFS betrieben werden, da lediglich IPFS auf der Client-Seite. Ein eventueller weiterer Lösungsansatz wäre eine externe Nutzung, z.B. in Kombination mit GitHub (dezentrale Ausführung), z.B. auch DockerHub.

Summary from the round table (2/2)

- Weiterführende Diskussion um bidirektionale oder unidirektionale Links sinnvoll.
- Durch IPFS/Filecoin wäre auch ein Umdenken in dem Punkt der Finanzierung des Storage möglich, indem man den Ansatz der Incentivierung verfolgt. Wie könnte man eine solche Incentivierung umsetzen und in die Praxis bringen. -> Vorgehensmodell
- Sehr günstiger Storage, da Überbietungen stattfinden, im Gegenzug „schnelles“ Geld für die Storage-Provider.
- Diskussion über den „Kurswert“ im Vergleich zu Bitcoin. Grundlage Storage-Contracts. Speicherung für „immer“ wird ein Storage-Provider nicht eingehen. Es gibt auch „Bestrafungsmechanismen“ für die Storage-Provider, wenn dieser die Vereinbarung nicht einhält. Ggf. würden zukünftige Storage-Contracts preistechnisch angepasst werden.
- IPFS ist nicht an Filecoin gekoppelt; es wären auch weitere Möglichkeiten gegeben, muss diesen Organisatorischen, sprich die Storage-Contracts zu verwalten und zu managen.

Questions?

If you have any questions, please feel free to contact Bela Gipp from the University of Göttingen (www.gipplab.org/gipp)
or the following Organizers

- Marco Beck (beck@gipplab.org)
- Cornelius Ihle (ihle@gipplab.org)
- Moritz Schubotz (schubotz@uni-wuppertal.de)

References and Links

- Concept for Setting up an LTA Working Group in the NFDI Section "Common Infrastructures“ (https://zenodo.org/record/6451456#.Yq9-WP5ByUm)