Decentralized web tools for scientific publishing

Prof. Dr. Philipp Koellinger

APE | 10 Jan 2023 Berlin

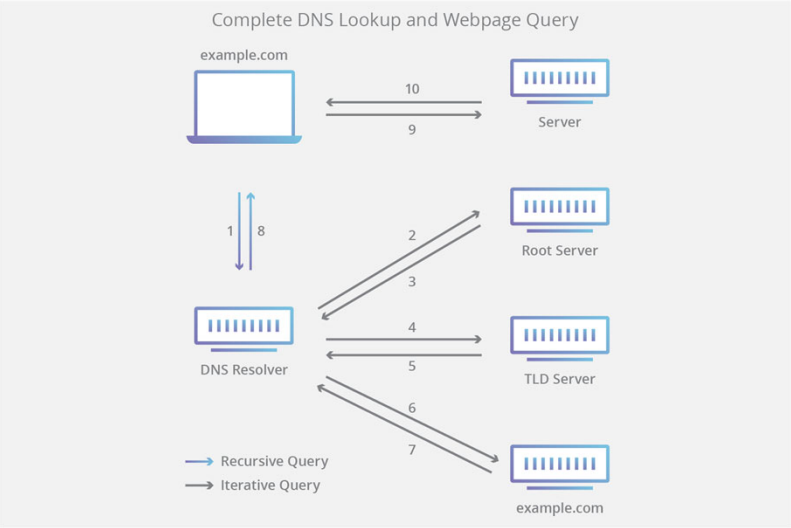
DeSci Foundation

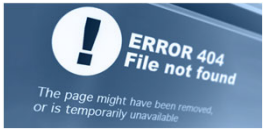
DeSci Labs

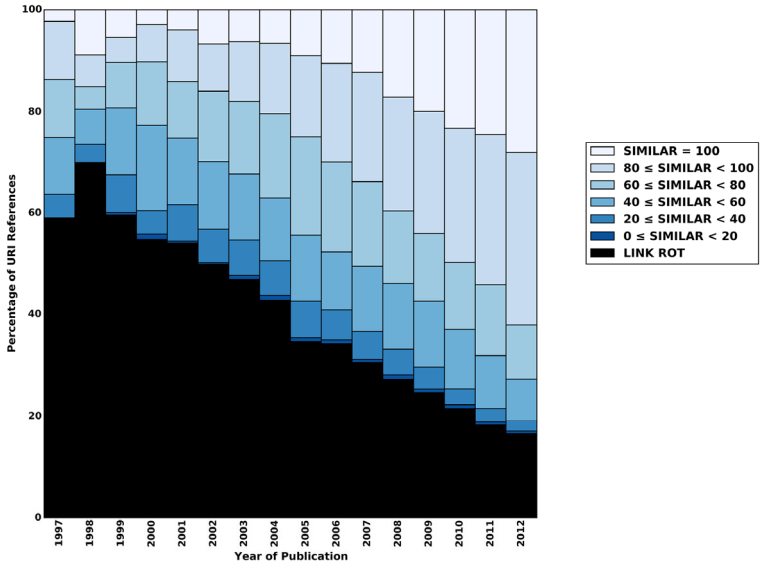
Decentralized web: Broad scope of innovations for scientific publishing

- Persistent identifiers based on hashes (PIDs)
  - Unique
  - Unbreakable
  - Permissionless (based on open-source software rather than a central authority)
  - New forms of citations (e.g. citations as function calls, interoperability, very fine-grained citations)
- Content-addressed data storage
  - More reliable (no link rot or content drift)
  - Cheaper (competitive marketplace for data storage, no manual updating required)
- Compute-over-data
  - Compute where the data is stored (efficient, cheap)
  - Verifiable and reusable compute outputs (trustworthy, efficient, interoperable)
- New tools for
  - Peer-review (e.g. DAOs, verifiable badges)
  - Content curation (e.g. Gateways)
  - Incentive design (e.g. rewarding referees for fast, high-quality reviews)
- New business models for scientific publishing

Identifiers in the current Internet – URIs







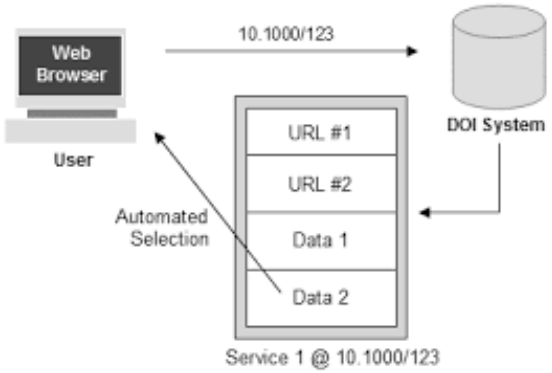
- URLs point to where content is stored, not what the content is
- Link rot (file moved or deleted, 404 error)
- Content drift (content changes over time)
- No version control

- URI citations with link rot or content drift
  by year of publication, Elsevier corpus (Jones et al. 2016)
- A threat to the integrity and value of the scientific record

Sources:

Jones, S.M., et. al. (2016). Scholarly context adrift: Three out of four URI References Lead to Changed Content. *PLoS ONE* 11(12): e0167475.

Identifiers in the current Internet – DOIs



- **DOIs do not correctly resolve to their target resource in ~50% of all cases** (Klein & Balakireva 2020)
- Different results for same DOI depending on request method and network environment
- DOIs are matched to URLs in a database
  → Lots of manual updating work for publishers
- Costly, inefficient system for publishers
- DOIs are neither persistent nor unique identifiers

Sources:
Klein, M., Balakireva, L. (2020). On the Persistence of Persistent Identifiers of the Scholarly Web. In: Hall, M., Merčun, T., Risse, T., Duchateau, F. (eds) *Digital Libraries for Open Knowledge*. TPDL 2020. Lecture Notes in Computer Science, vol. 12246. Springer.

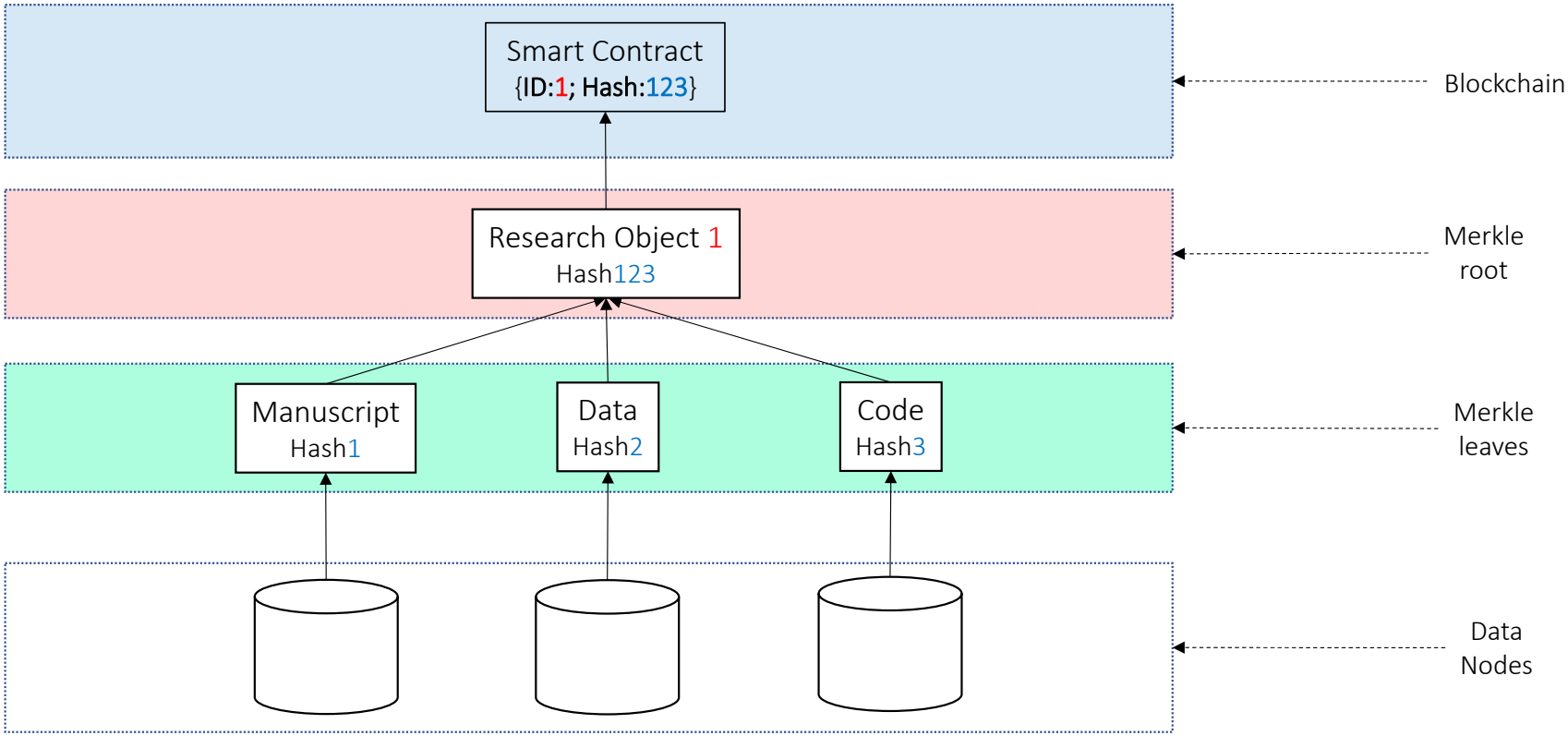
Content addressing based on hashes

- A cryptographic hash function converts a string of arbitrary length into string of fixed length
  - One-way mathematical function
    - E.g., the SHA-256 algorithm creates a 64 hexadecimal string for any input
  - Changing *anything* in the input (i.e. a word, pixel, comma) will yield a different hash
  - Hashes are *unique*
    - E.g., SHA-256 allows creating 10^{77} different hashes – billions of times more than the number of atoms on Earth
  - For example, SHA-256 hashes:
    - "Brazil will win the Fifa Worldcup 2026" → 157a222e95daa553283bcbdf73f124fc6119a0fbe285d2a7f40fa39ea8cc751f
    - "Argentina will win the Fifa Worldcup 2026" → a73879c974dfe4d3431897b26d64559bdbb5a81e8f7bc2504f19c6e7d75fc218

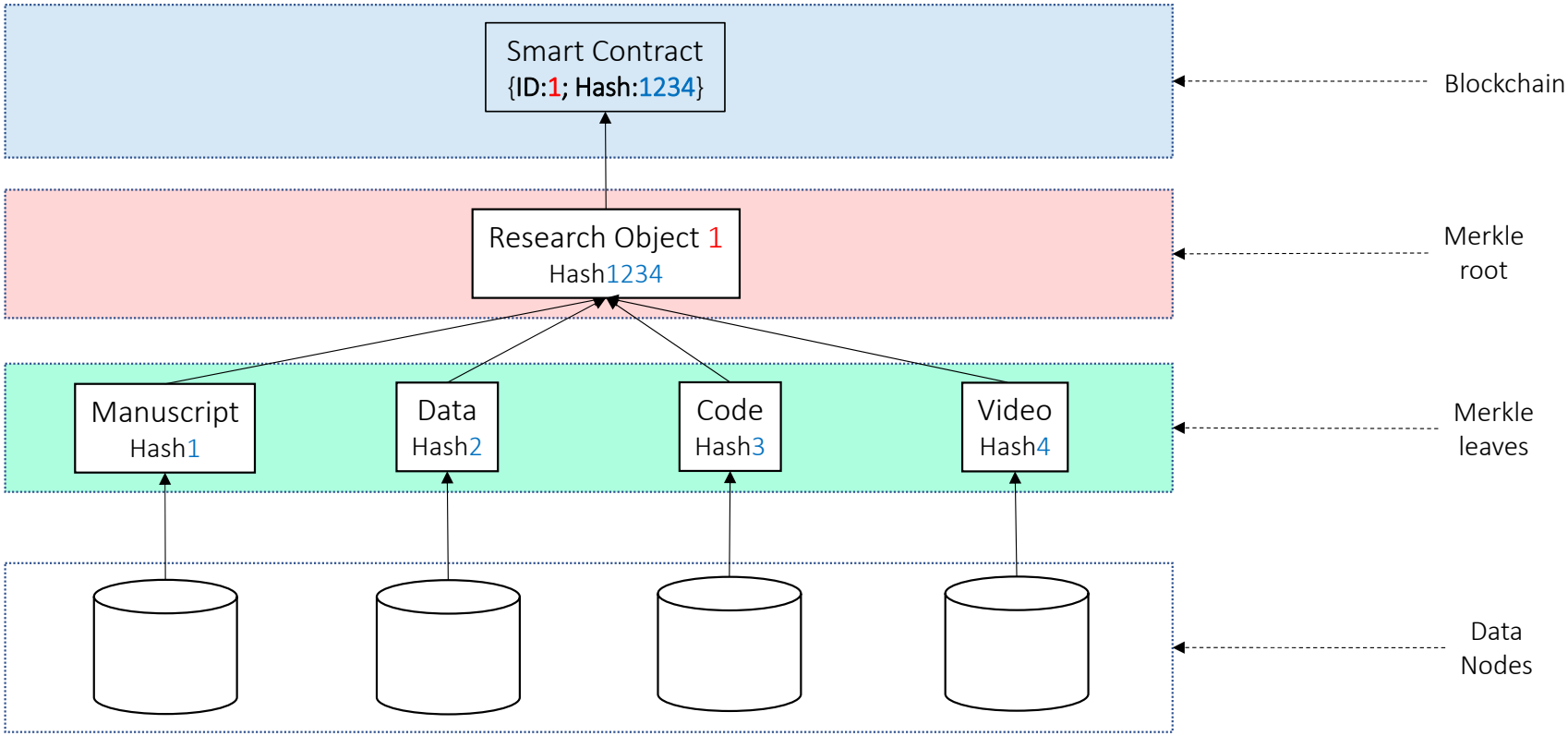
- Content addressing based on cryptographic hashes is immune to content drift and link rot!

Sources:
https://xorbin.com/tools/sha256-hash-calculator

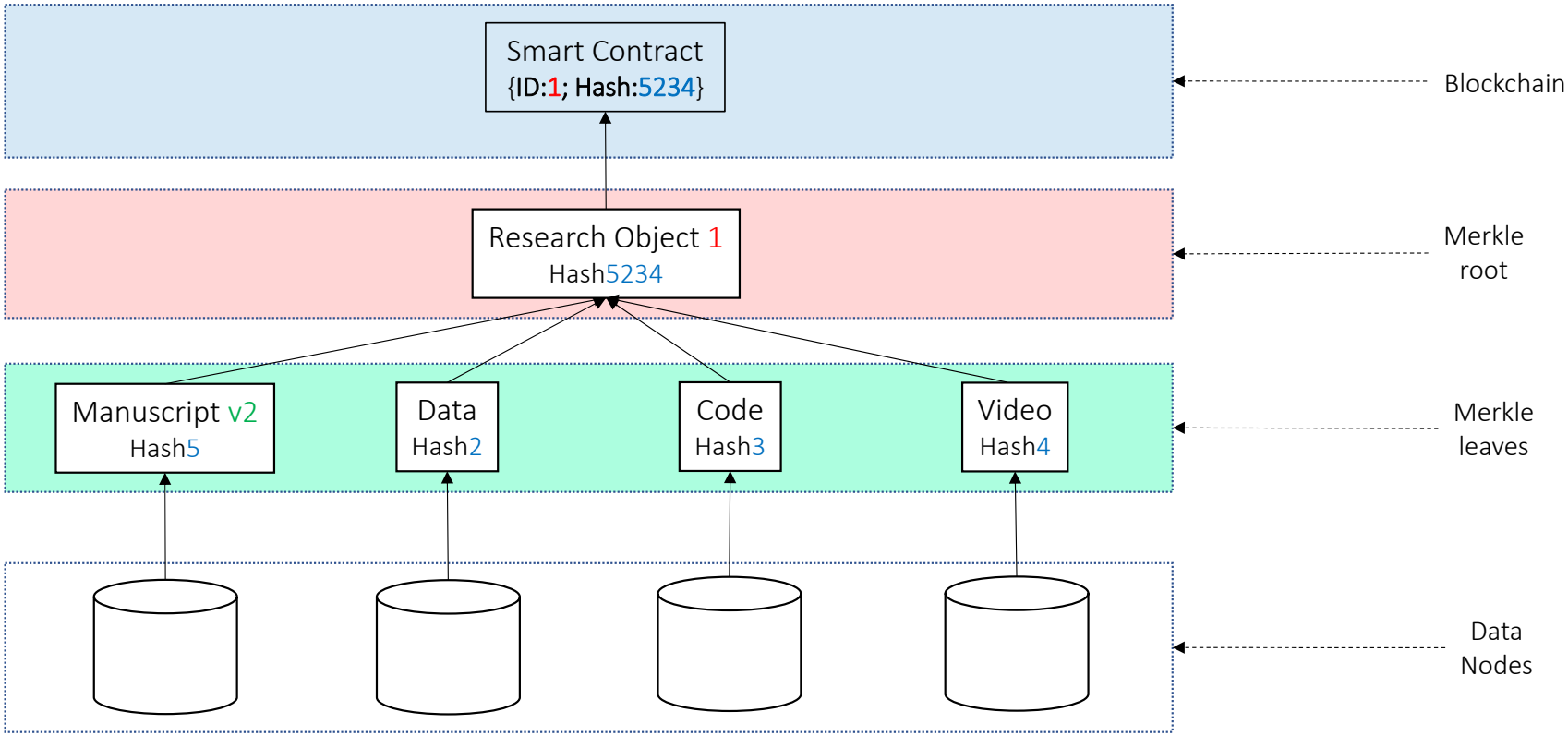
Rich research objects with hash-PIDs, indexed on a blockchain



Adding a new component to the research object

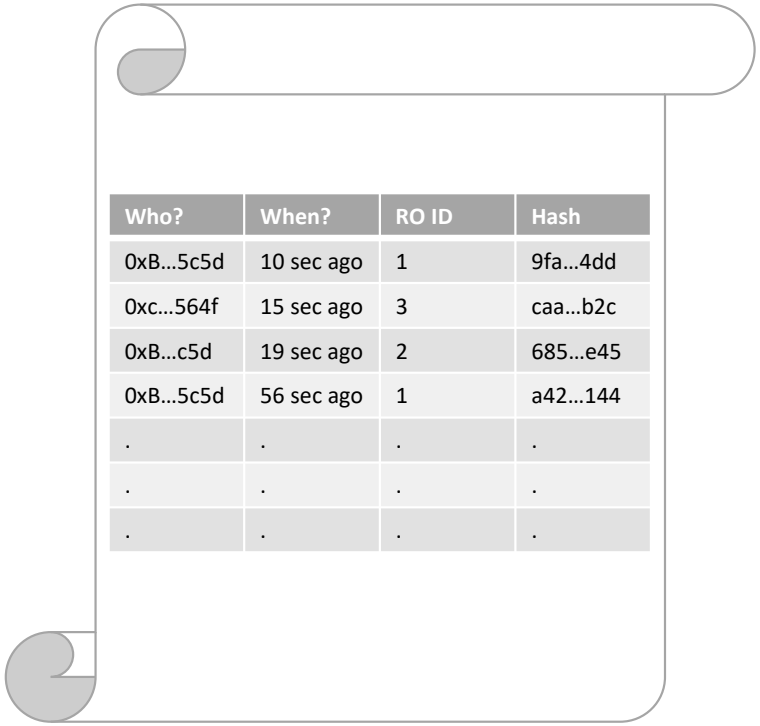


Updating a component of the research object



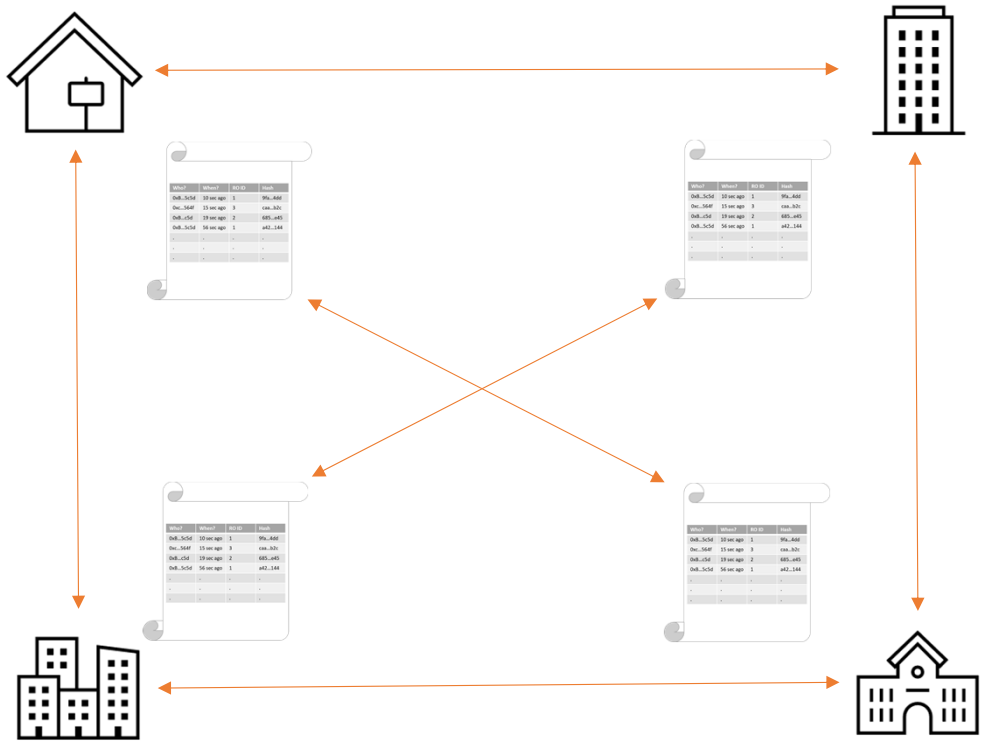
→ The decentralized web allows us to upgrade the scientific record from static manuscripts *without* persistent IDs *or* version control to rich, dynamic, interoperable research objects *with* persistent IDs *and* version control!

A public registry of research objects on a blockchain



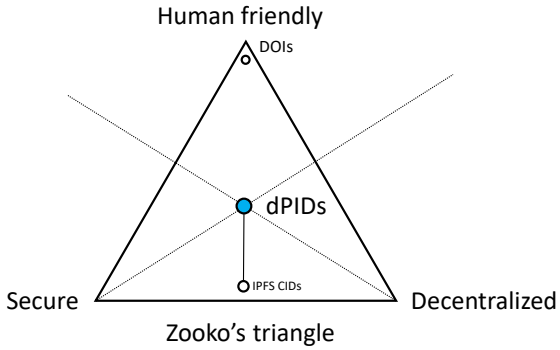
Who?	When?	RO ID	Hash
0xB…5c5d	10 sec ago	1	9fa…4dd
0xc…564f	15 sec ago	3	caa…b2c
0xB…c5d	19 sec ago	2	685…e45
0xB…5c5d	56 sec ago	1	a42…144
.	.	.	.
.	.	.	.
.	.	.	.

…distributed across many servers



Persistent identifiers

- PIDs that address the entire linked data structure of a research object directory
- Structure: {Resolver}/{PID}/{version identifier OR CID}/{Component index}/{Component suffix}
- Examples (all the same):
  - Long format:
    - dpid.org/42/bafybeigdyrzt5sfp7udm7hu76uh7y26nf3efuylqabf3oclgtqy55fbzdi/1/measurements.csv
  - Short, human friendly:
    - dpid.org/42/0/1/measurements.csv
    - dpid.org/42/v1/data/measurements.csv
  - Interoperable:
    - http_import(dpid.org/42/v1/data/measurements.csv)



Granular, version-controlled citations and citations as function calls

- Granular, version-controlled citations
  - Reference:
    - a particular sentence from a paper,
    - a line of code, or
    - a specific part of a dataset.

- Citations as function calls
  - One research object calls and executes a piece of code from a different research object on chain
  - Run code from your own research object on data from a different research object
  - Video demonstration: YouTube “DeSci Nodes Product Update 2022-08”
    - https://www.youtube.com/watch?v=VgvzuHf9j-s&t=166s

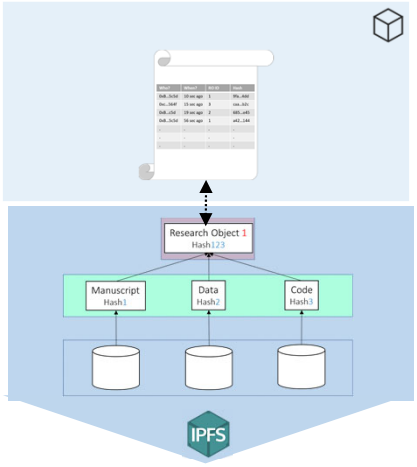
→ Enabling interoperable research objects

An open protocol for FAIR, interoperable research objects…

Permission management and PID registry
"What PID/hash pair does this PID URI correspond to?"
"Who has the right to write a version update?"

Version indexing
"What version CID hash corresponds to this PID?

Digital object indexing
"What are the digital objects linked to this version CID hash?"



REGISTRY LAYER

Research Object PID
Ledger entry

DATA LAYER

Content-addressed decentralized storage network
3 continent, 5 countries, different storage providers
(multiple copies keep things safe)

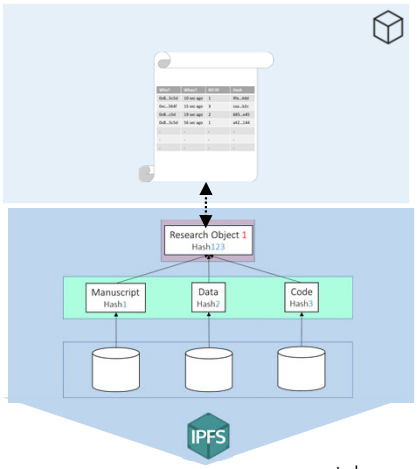
…with compute integration…

Permission management and PID registry
"What PID/hash pair does this PID URI correspond to?"
"Who has the right to write a version update?"

Version indexing
"What version CID hash corresponds to this PID?

Digital object indexing
"What are the digital objects linked to this version CID hash?"

Compute integration
"Is this work reproducible? How can I easily run an analysis on that data?"



REGISTRY LAYER

Research Object PID
Ledger entry

DATA LAYER

Content-addressed decentralized storage network
3 continent, 5 countries, different storage providers
(multiple copies keep things safe)

Compute Job + Data Hash = Job Output Hash

DeSci Foundation

@DesciLabs
@DesciFoundation
@PKoellinger

…and open index and shortened, human adapted PIDs…

"Report real time data analytics on research objects"

Open Index

INDEXING LAYER
GraphQL indexer
Easy and fast to analyze the data

"Return the data and metadata linked to this PID over HTTP"

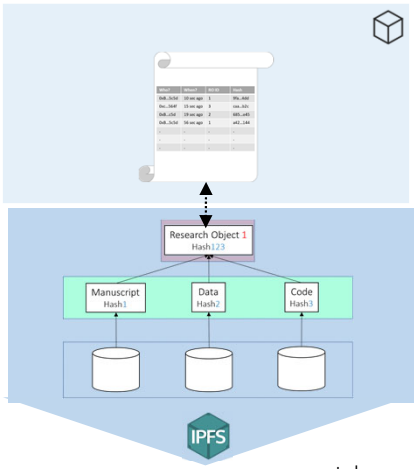
dpid.org

DNS RESOLVER
Shortened, human adapted PIDs

Permission management and PID registry
"What PID/hash pair does this PID URI correspond to?"
"Who has the right to write a version update?"

Version indexing
"What version CID hash corresponds to this PID?

Digital object indexing
"What are the digital objects linked to this version CID hash?"



REGISTRY LAYER
Research Object PID
Ledger entry

DATA LAYER
Content-addressed decentralized storage network
3 continent, 5 countries, different storage providers
(multiple copies keep things safe)

Compute integration
"Is this work reproducible? How can I easily run an analysis on that data?"

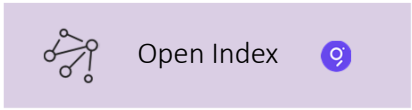
Compute Job + Data Hash = Job Output Hash

DeSci Foundation

@DesciLabs
@DesciFoundation
@PKoellinger

…and Gateways that serve as curation tools

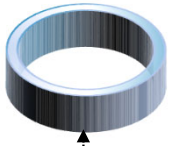
"Report real time data analytics on research objects"



INDEXING LAYER
GraphQL indexer
Easy and fast to analyze the data

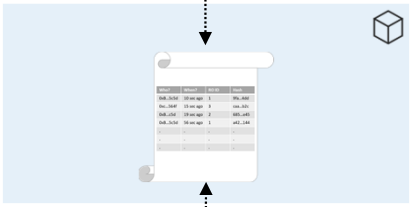
"Return the data and metadata linked to this PID over HTTP"

DNS RESOLVER
Shortened, human adapted PIDs



Gateways
(e.g. journals, preprint platforms, open access libraries, funders)

Open read/Open write

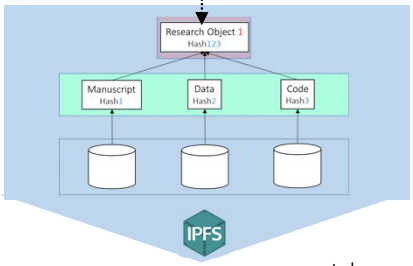


REGISTRY LAYER

Research Object PID
Ledger entry

Permission management and PID registry
"What PID/hash pair does this PID URI correspond to?"
"Who has the right to write a version update?"

Version indexing
"What version CID hash corresponds to this PID?



DATA LAYER

Content-addressed decentralized storage network
3 continent, 5 countries, different storage providers
(multiple copies keep things safe)

Digital object indexing
"What are the digital objects linked to this version CID hash?"

Compute integration
"Is this work reproducible? How can I easily run an analysis on that data?"

Compute Job + Data Hash = Job Output Hash

DeSci Foundation

@DesciLabs
@DesciFoundation
@PKoellinger

Summary – An upgrade to the scientific record

- A complete open-science publication protocol
  - Manuscripts, data, code, videos, metadata
  - Persistent identifiers
  - Version control
  - Enabling FAIR compliance
  - Petabyte-scale research objects possible
  - Compute-over-data
  - Enabling easier reproducibility
- No more link rot or content drift
- Substantial cost savings
  - Cheap storage
  - Low maintenance (e.g. no more manual updating of DOI entries)
  - Data integrity
  - Cheaper data cataloguing

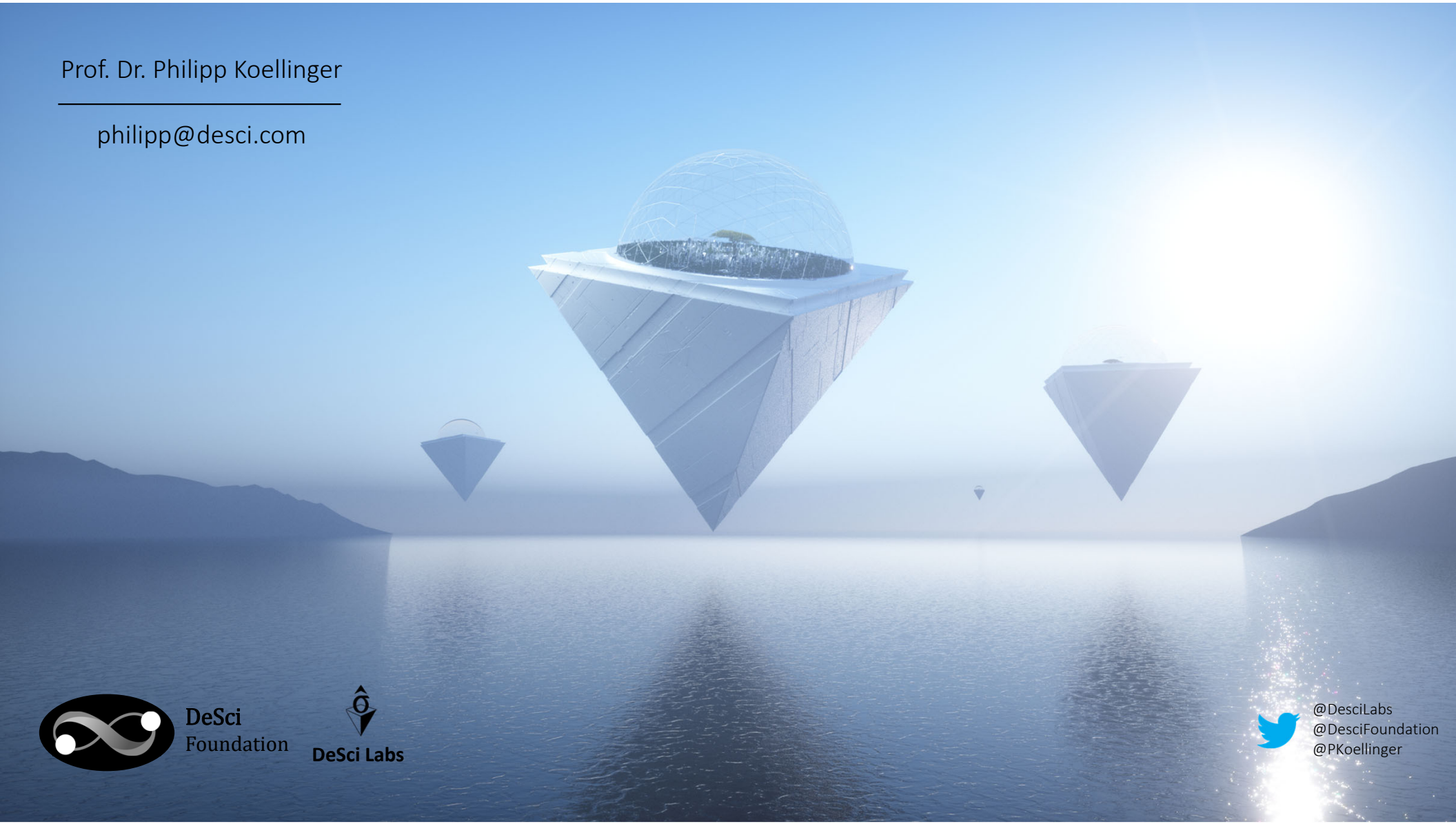
Discussion

- Perhaps publishers could outsource a substantial part of their IT to such a system?
  - Journals as Gateways
  - Gateways could charge APCs or submission fees

- Why is this NOT going to work?

Prof. Dr. Philipp Koellinger

philipp@desci.com



DeSci Foundation

DeSci Labs

@DesciLabs
@DesciFoundation
@PKoellinger