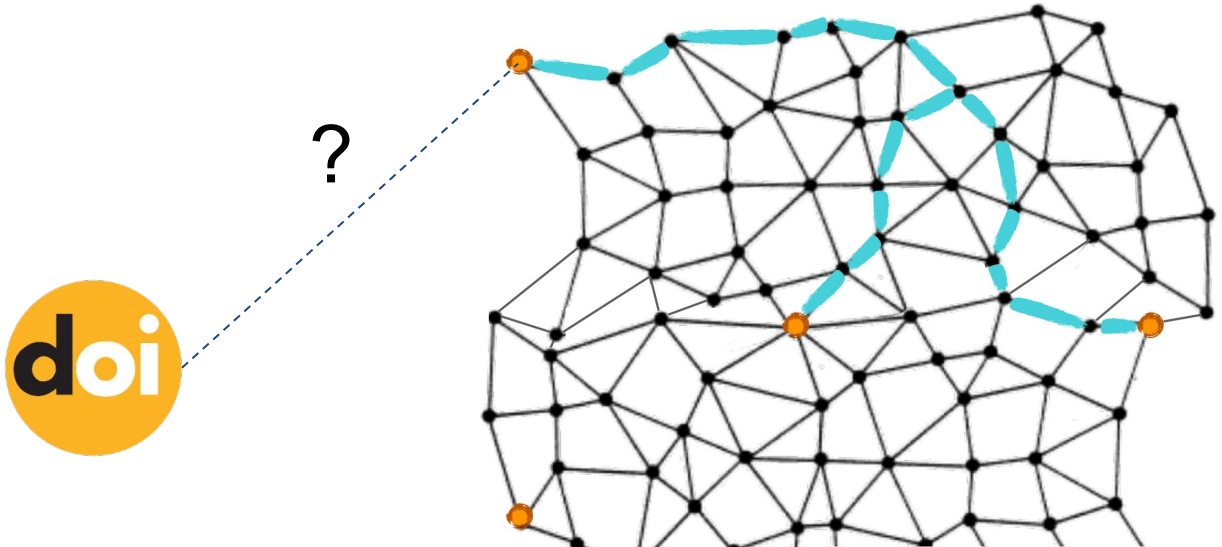
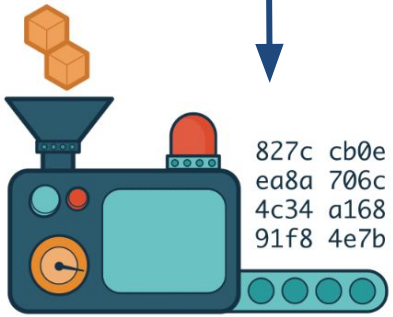
# p(id)-2-p(id):
# bridging PIDs to the p2p, Content-Addressed Web?

Eoghan Ó Carragáin
PIDapalooza
2019-01-23



**LOCATION-ADDRESSING**

http://site.com/data/pids.pdf

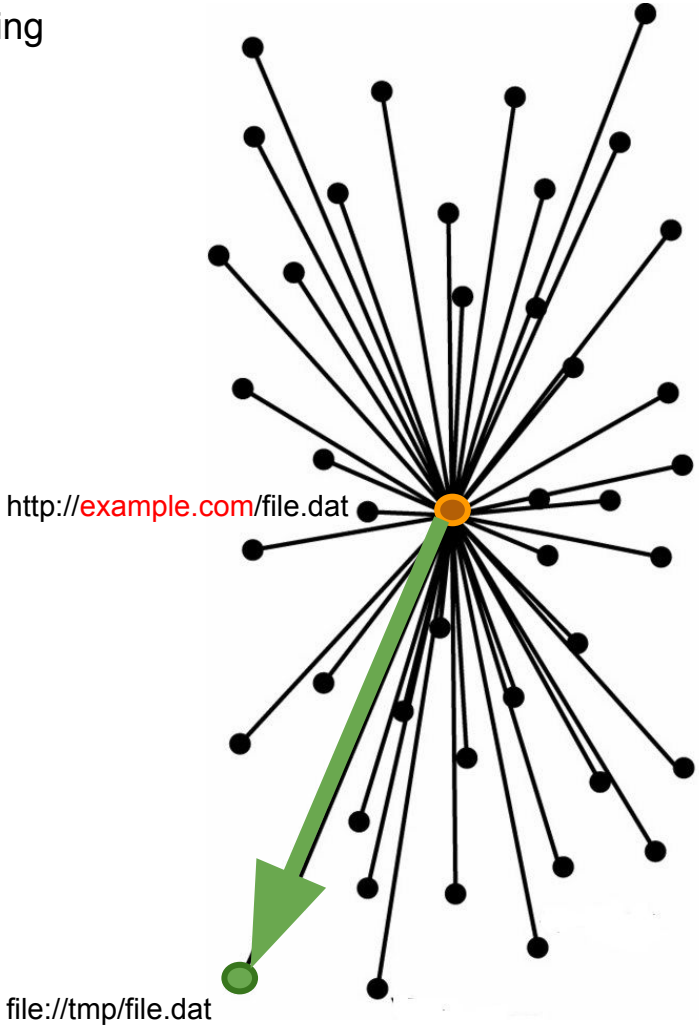



ipfs://zdj7WjqNrjReTcEveRh/pids.pdf

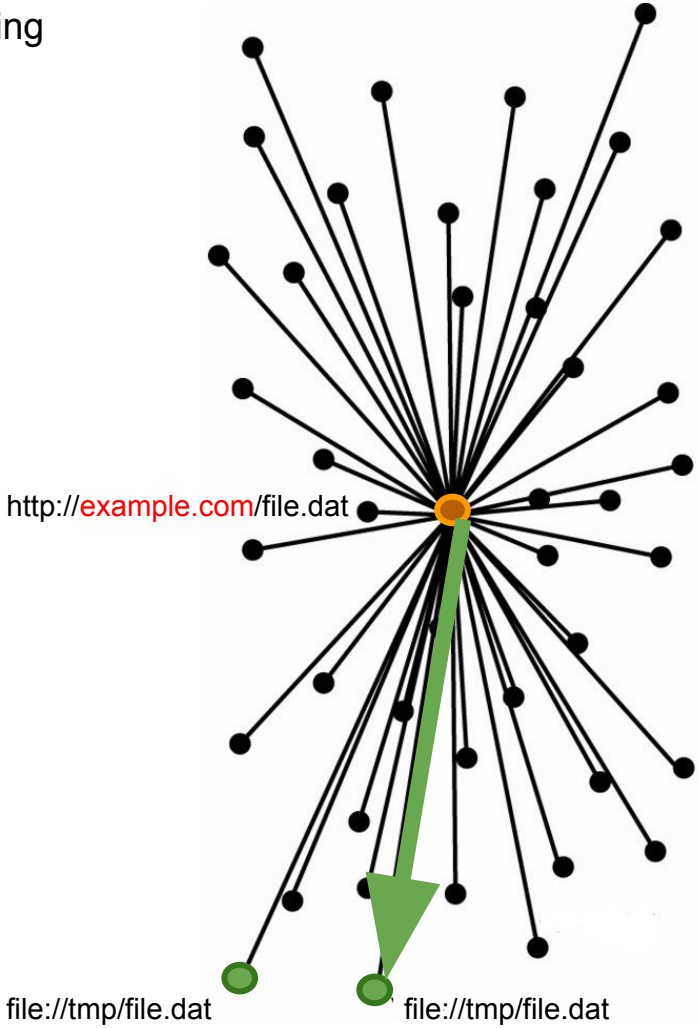
**CONTENT-ADDRESSING**

Image: @protocollabs

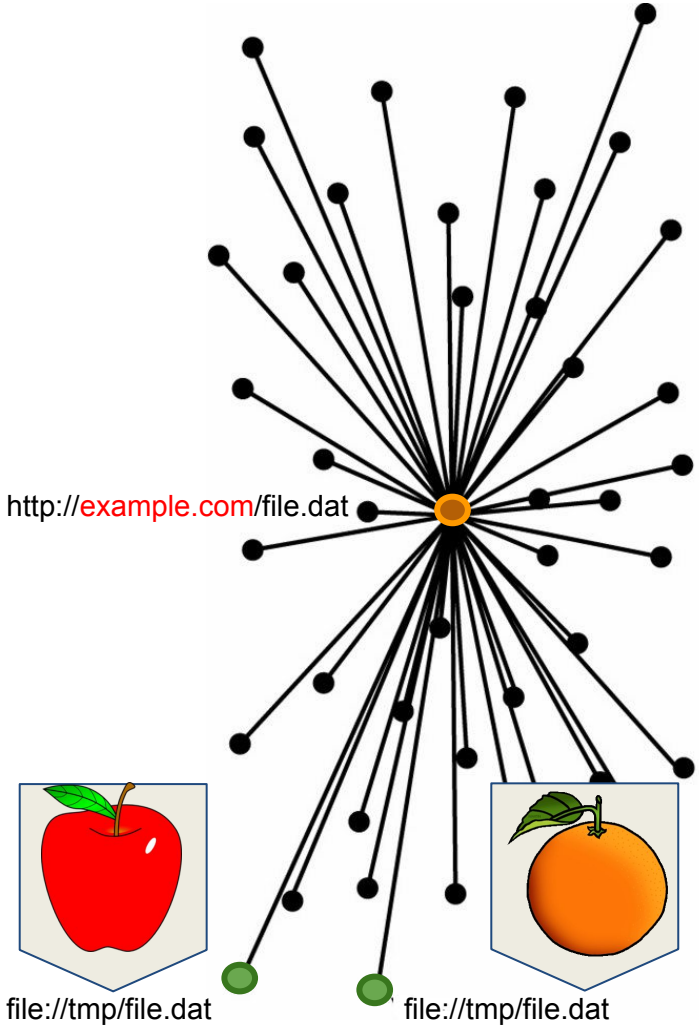
# Centralisation through location addressing



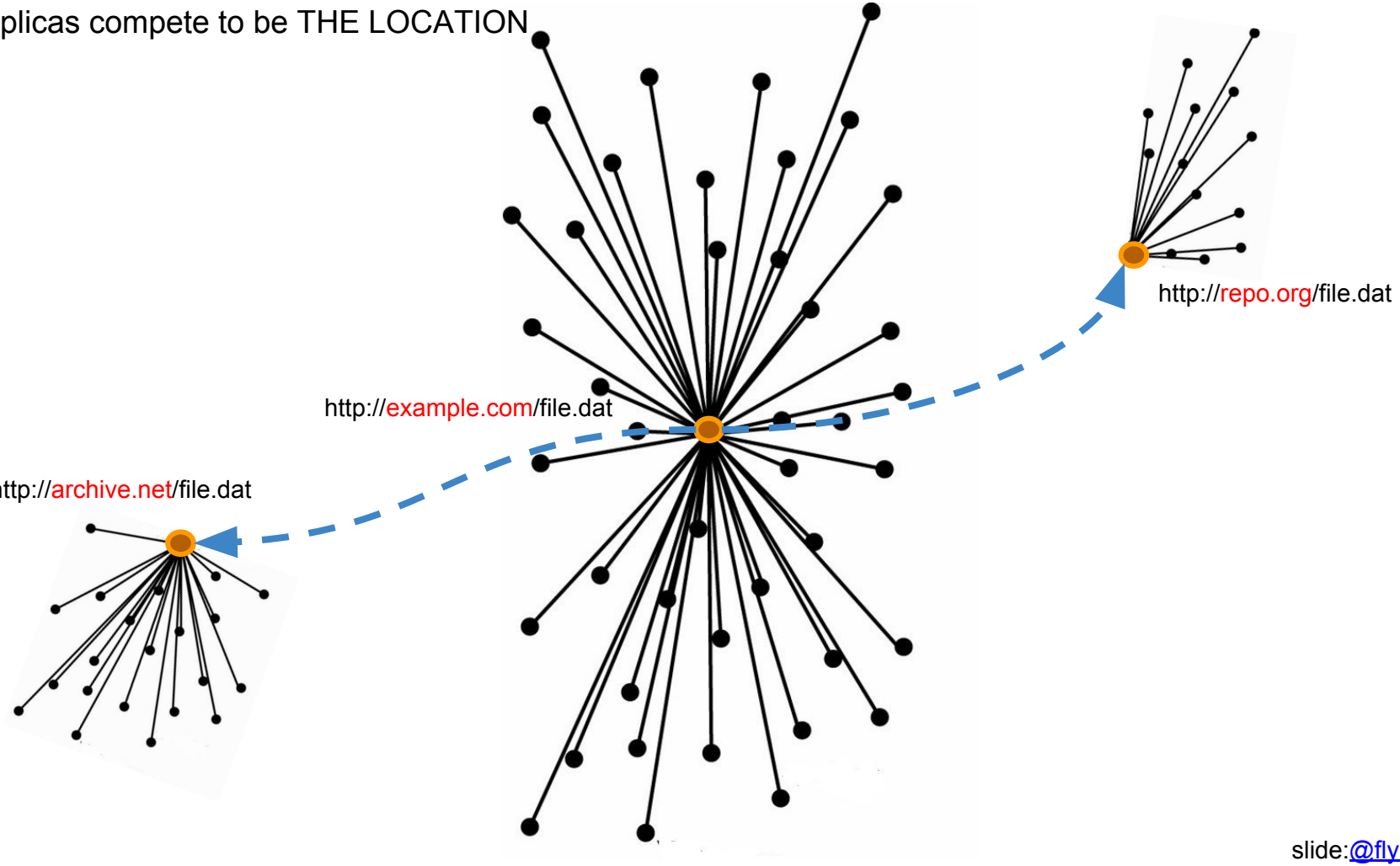
# Centralisation through location addressing



Web references are mutable & “content negotiable” by design

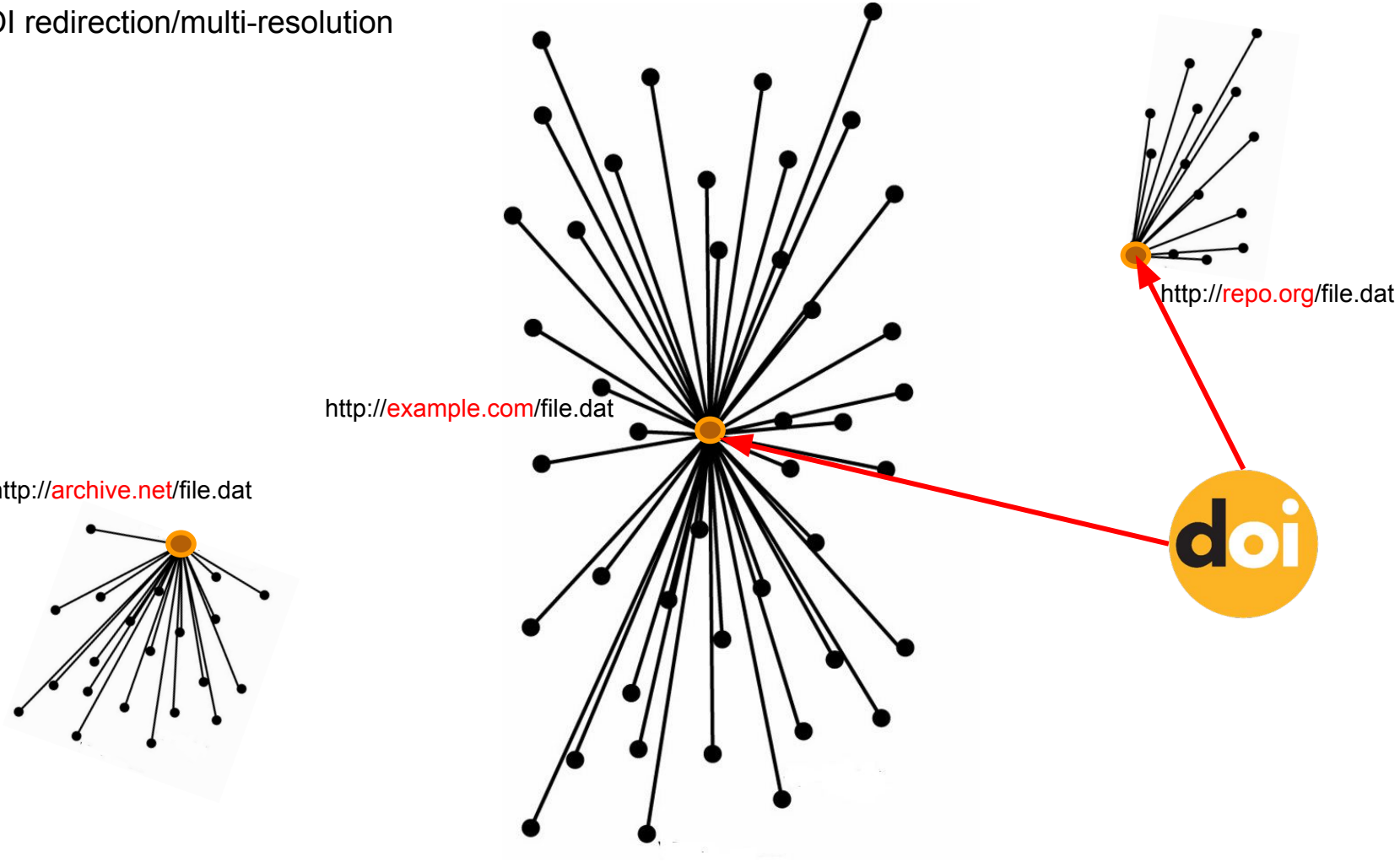


# Replicas compete to be THE LOCATION

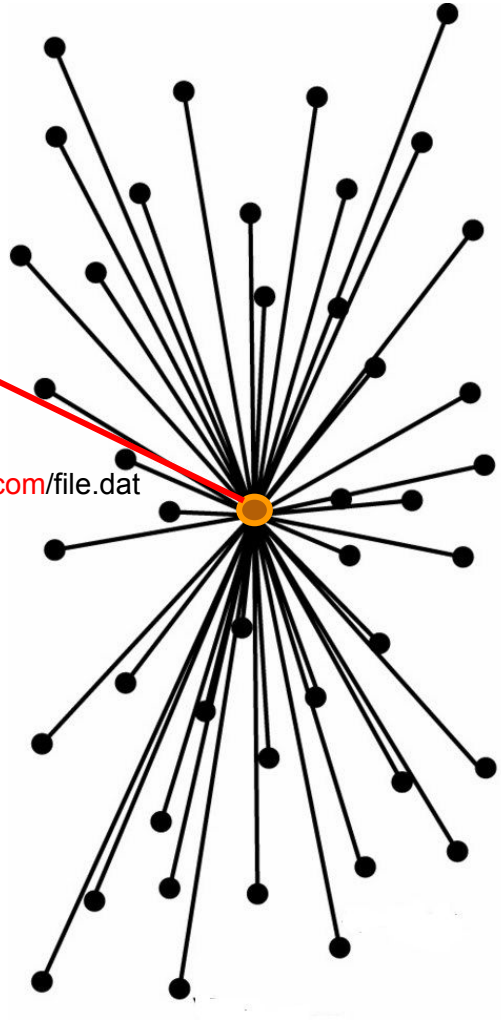



slide:@flyingzumwalt

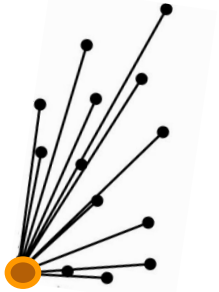
# DOI redirection/multi-resolution



Memento: find replica by time-based content-negotiation

http://example.com/file.dat

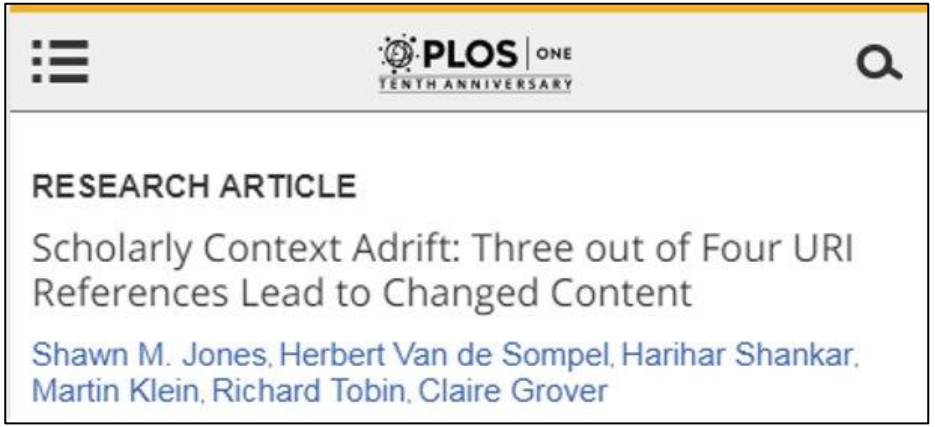


http://repo.org/file.dat

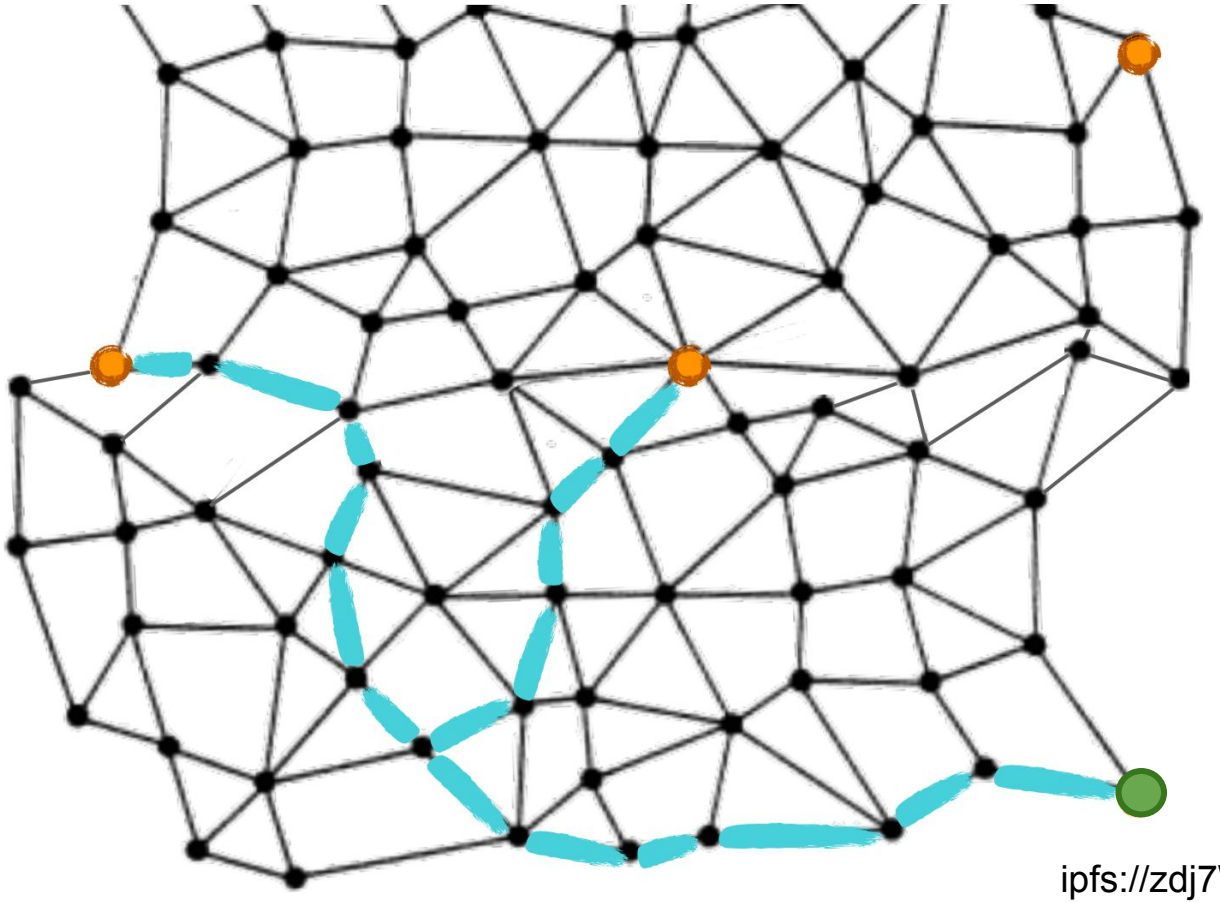
http://archive.net/file.dat

# Old favourites: “Link rot” and “content drift“


https://doi.org/10.1371/journal.pone.0115253




https://doi.org/10.1371/journal.pone.0167475



ipfs://zdj7WjqNrjReTcEveRhgcsXsJvSGwLxJ7js1R7ZCzNaQSKuTh

image:[@flyingzumwalt]

You Retweeted

**Pieter J. Van Garderen** @pjvangarderen · Apr 11

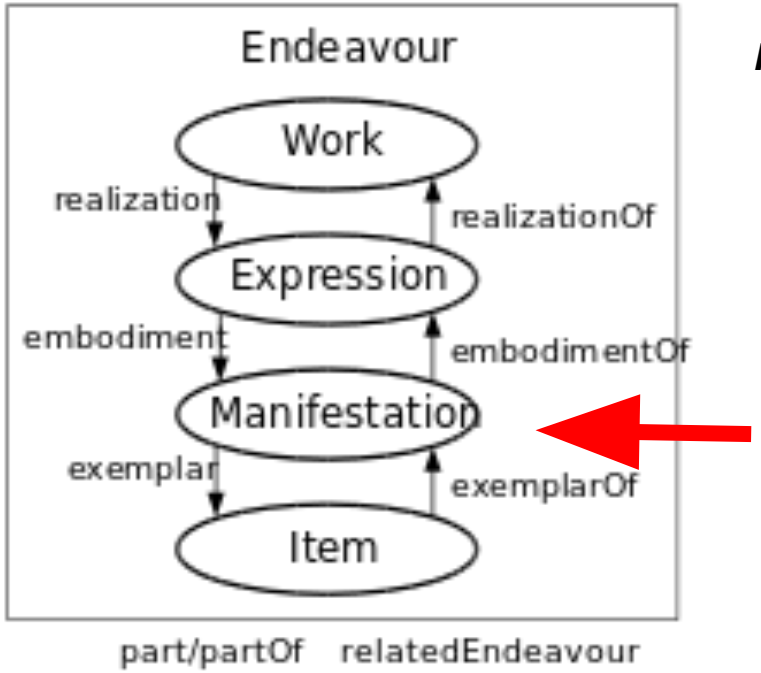
In my 20 years experience in the #digipres domain I have read a fair share of complex theory, principles, etc.. In practice, I am always able to simplify 'things' and group them under three core questions: 1) can I find it? 2) can I use it? 3) can I trust it?

3 19 29

# Some research data use-cases for immutable, predictable and verifiable addressing

# URIs != URLs!

Web Resources and DOIs can identify (often un-hashable) physical, digital, and abstract *things*




***For now*** let's focus on static/versioned digital content, e.g.:

- a specified/canonical "representation" of an "information resource" in web terms
- a "digital creationStructuralType" in the DOI Data Dictionary terms
- a specific "manifestation" in FRBR terms
- a "payload" in BagIt terms
- "Identifiers for Digital Objects" rather than "Digital Identifiers of Objects"

Software Heritage
PRESERVING TECHNICAL KNOWLEDGE

# Data citation and versioning

“The demand for reproducibility of research results is growing. [There is need] to **reference the exact version of the data** that was used to underpin the research findings, and/or was used to generate higher level products. ”

RDA
RESEARCH DATA ALLIANCE

**Data Citation WG**

**Data Versioning WG**

# Data citation and versioning


```
{
  "relatedIdentifier": "10.5281/zenodo.580337",
  "relatedIdentifierType": "DOI",
  "relationType": "HasVersion"
}
```

## Versions

| | | |
|---|---|---|
| Version 2.2 | 10.5281/zenodo.580337 | May 16, 2017 |
| Version 2.1.3 | 10.5281/zenodo.48270 | Mar 24, 2016 |
| Version 2.1.2 | 10.5281/zenodo.48068 | Mar 21, 2016 |



Any C.U.D. operation on files triggers a new version.


| Version(s) | 1 | 2 |
|---|---|---|
| REVISED Version 2 published 18 Jan 2019 | | |
| Version 1 published 05 Nov 2018 | ? read report | ✓ read report |

# The need to verify the exact content

**Table 1: Mechanism implementation in common systems of identifiers**

| Mech. / System | Handle | DOI | Ark | PURL | VDOI |
|---|---|---|---|---|---|
| Generation | Yes | Yes | Yes | Yes | Yes |
| Assignment | Yes | Yes | Yes | Yes | Yes |
| Verification | N.A. | N.A. | N.A. | N.A. | Yes |
| Retrieval | Yes | Yes | Yes | Yes | Yes |
| Reverse Lookup | N.A. | N.A. | N.A. | N.A. | N.A. |
| Description | Yes | Yes | Yes | N.A. | Yes |


https://hal.archives-ouvertes.fr/hal-01865790

# Static, portable “Data Packages”

researchobject.org

BagIt

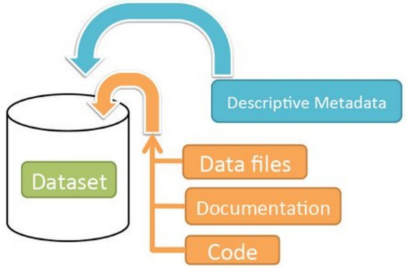



image: @OA_RHUL

FRICTIONLESS DATA
STANDARDS AND TOOLING

DataONE
Data Observation Network for Earth

Portable Encapsulated Project


DwC Archive

o2r

ReproZip

“Make **Data Crate** again!”

# Direct linking of PIDs to downloadable content (with hashes)

minid


PID Kernel Information WG


Kernel record/metadata

Direct access to content associated with a DOI #2

Open mfenner opened this issue on Nov 30, 2017 · 23 comments

# Hashes in PIDs and URIs

Trusty URIs

http://example.org/r1.RAcbjcRIQozo2wBMq4WcCYkFAjRz0AX-Ux3PquZZrC68s

ICOS INTEGRATED CARBON OBSERVATION SYSTEM

hdl:11676/6T0zQII1VzJHDmJLSZU5s4qE

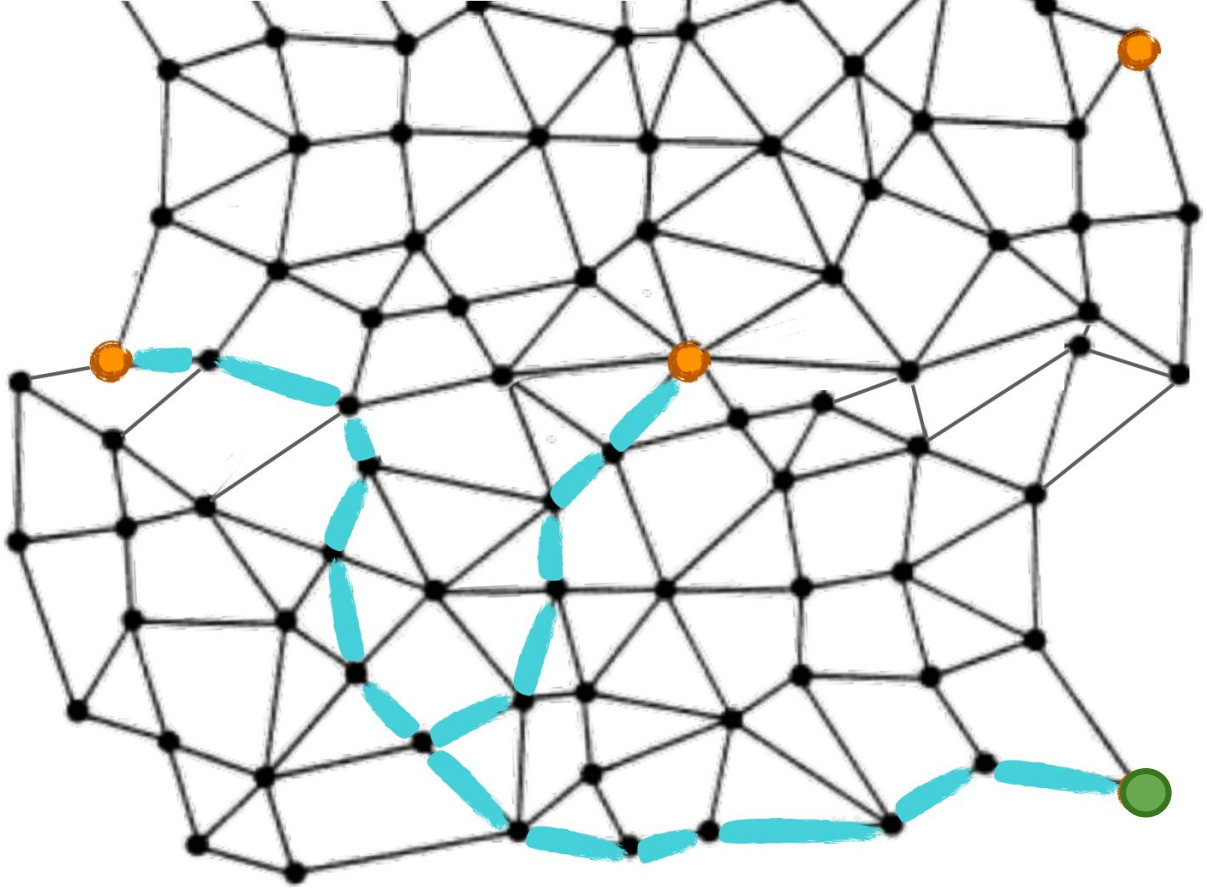
Software Heritage

PRESERVING TECHNICAL KNOWLEDGE

swh:1:cnt:94a9ed024d3859793618152ea559a168bbcbb5e2


Which brings us back to to content-addressing and IPFS

ipfs://zdj7WjqNrjReTcEveRhgcsXsJvSGwLxJ7js1R7ZCzNaQSKuTh



How does IPFS help with link rot?

- Anyone can ‘mint’ an IPFS identifier, i.e. relatively persistent “web-at-large” identifiers
- Data is available as long as any node on the network shares it
- Replication is trivial, verifiable and reinforces availability (LOCKSS)
- Clusters of nodes can coordinate to ‘pin’
- Persistence becomes participatory

How does IPFS help with content drift?

- Referenced data is immutable by design
- Integrity check is part of dereferencing
- Fine-grained access/citation of sub-resources (“range of verifiability”)
- Has an underlying data-model (IPLD) that can be used to express all sorts of relevant data structures: file-systems, git-like versioning, virtual aggregations across datasets (e.g. OAI-ORE)

# Great, so let's just use IPFS…?

Immutability != permanent/persistent availability

- Who coordinates 'nodes of last resort' (c.f. Keepers Registry)?
- Persistent availability of large amounts of research data = **Collective Action Problem** (see: 10.5334/kula.7)

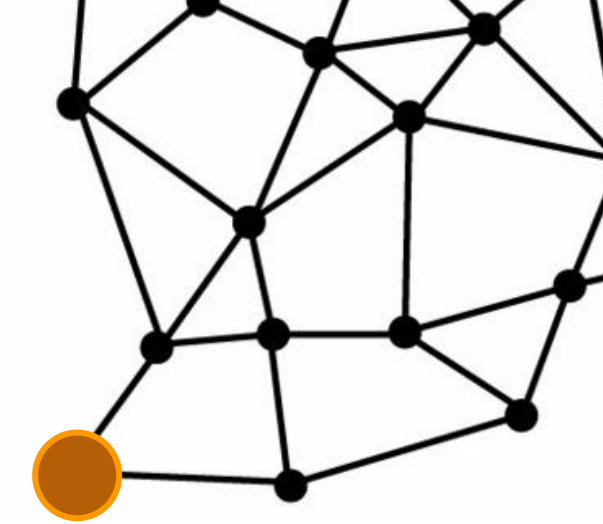
Inevitability of hash collisions (at some stage)

- Time-frame good enough for "web-at-large" and "intermediate, often transient, data products" (e.g. MINIDS)?
- For published, scholarly record, you'd need an indirection layer, to be able to update citations to point at new hashes (sound familiar?)

Maturity, adoption, stability, maintenance of (any) technology

- For long-term persistence, we need an indirection layer which allows upgrading between technology stacks and protocols (sound familiar?)

The challenge of persisting research data is ultimately social: people, organisation, communities, governance.



… but let's adopted use the technologies and network paradigms that fit that collective mission best!

# Answers?