ETH zürich

ETH Library

What you cite is what you get? Verifiable addressing of immutable, self-describing research data

Other Conference Item

Author(s):
Ó Carragáin, Eoghan

Publication date:
2019-09-13

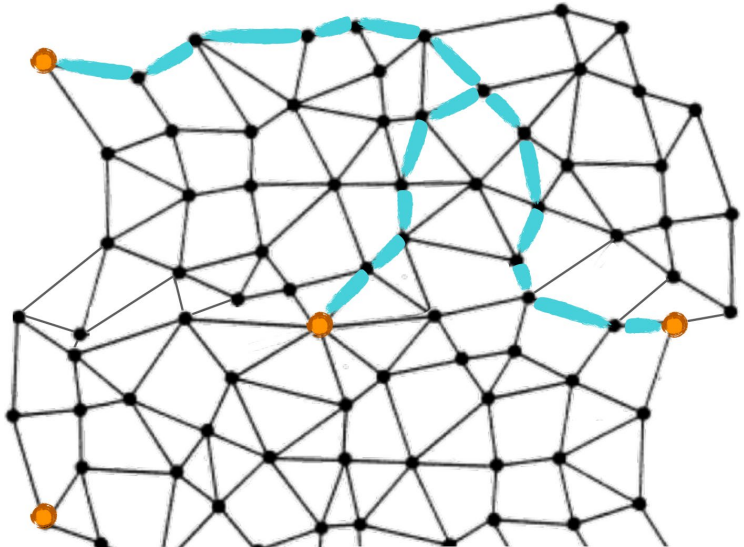
Permanent link:
https://doi.org/10.3929/ethz-b-000366549

Rights / license:
In Copyright - Non-Commercial Use Permitted

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

What you cite is what you get?
Verifiable addressing of immutable, self-describing research data

Eoghan Ó Carragáin
Persistent Identifiers in Research
ETH Zurich
2019-09-13



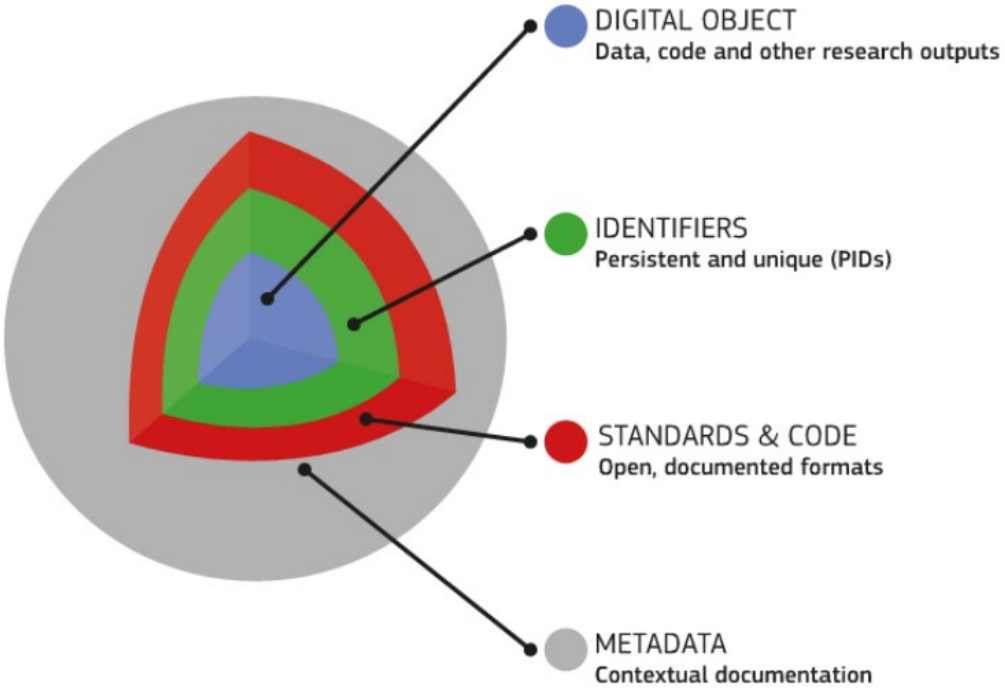
You Retweeted

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

Can I Use It? “FAIR Digital Objects”



Turning FAIR into Reality 10.2777/1524

Can I Use It? Portable, self-describing “Data Packages”

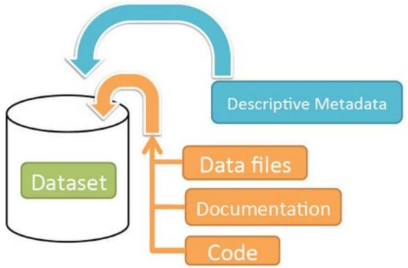


image: @OA_RHUL

BDBags

researchobject.org

FRICTIONLESS DATA
STANDARDS AND TOOLING

Psych-DS

Portable Encapsulated Project

DataONE
Data Observation Network for Earth

RO-Crate

o2r

ZIP

DwC Archive

ReproZip

Can I Trust it? Citation and versioning of **Research Data**

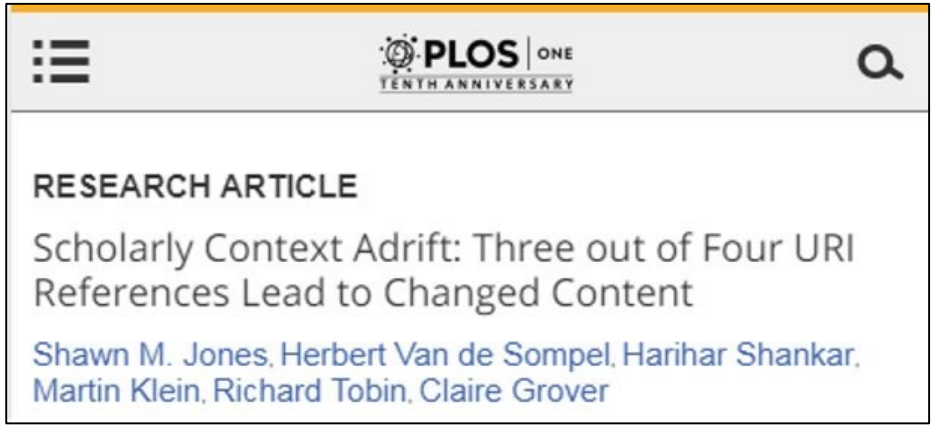
"The demand for **reproducibility** of research results is growing. [There is need] to **reference the exact version of the data** that was used to underpin the research findings, and/or was used to generate higher level products. "

Data Citation WG

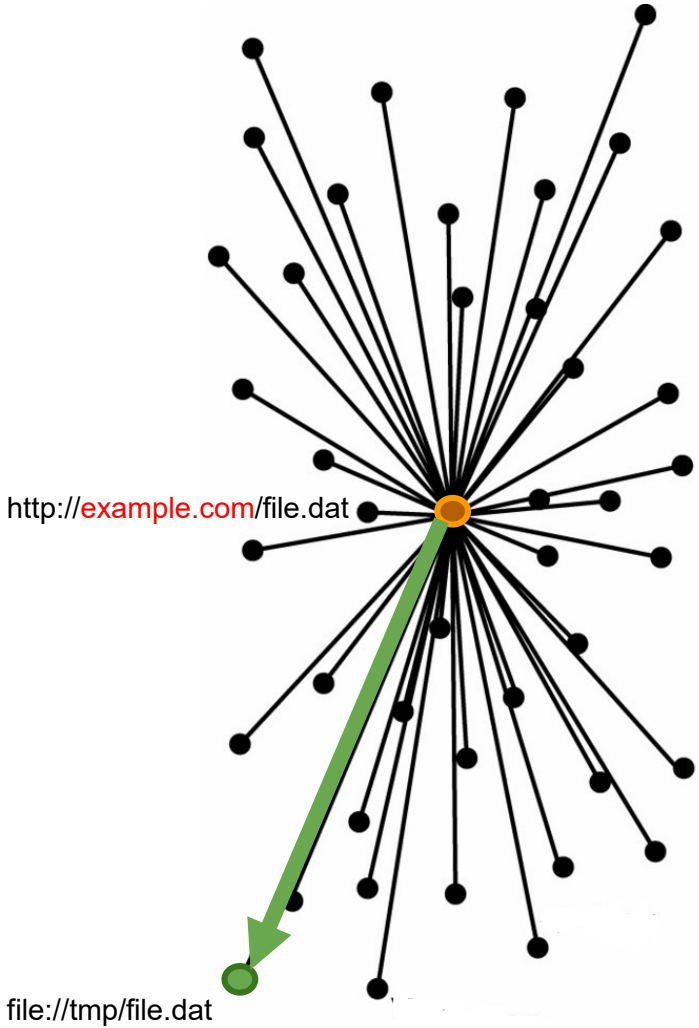
Data Versioning WG

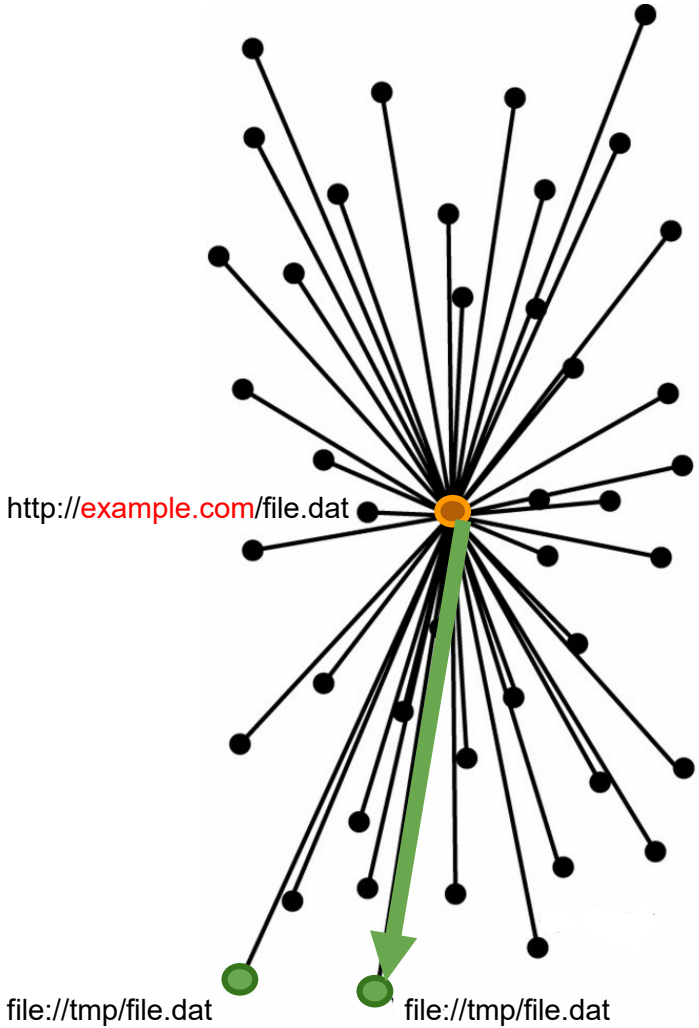
“Link rot” and “content drift”

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

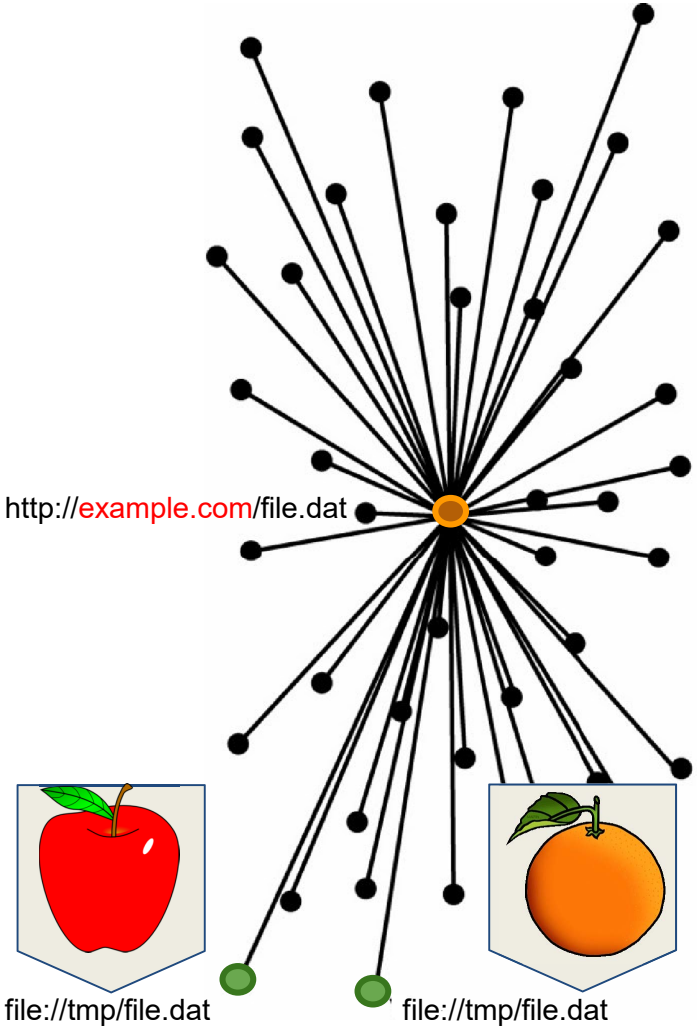


https://doi.org/10.1371/journal.pone.0167475

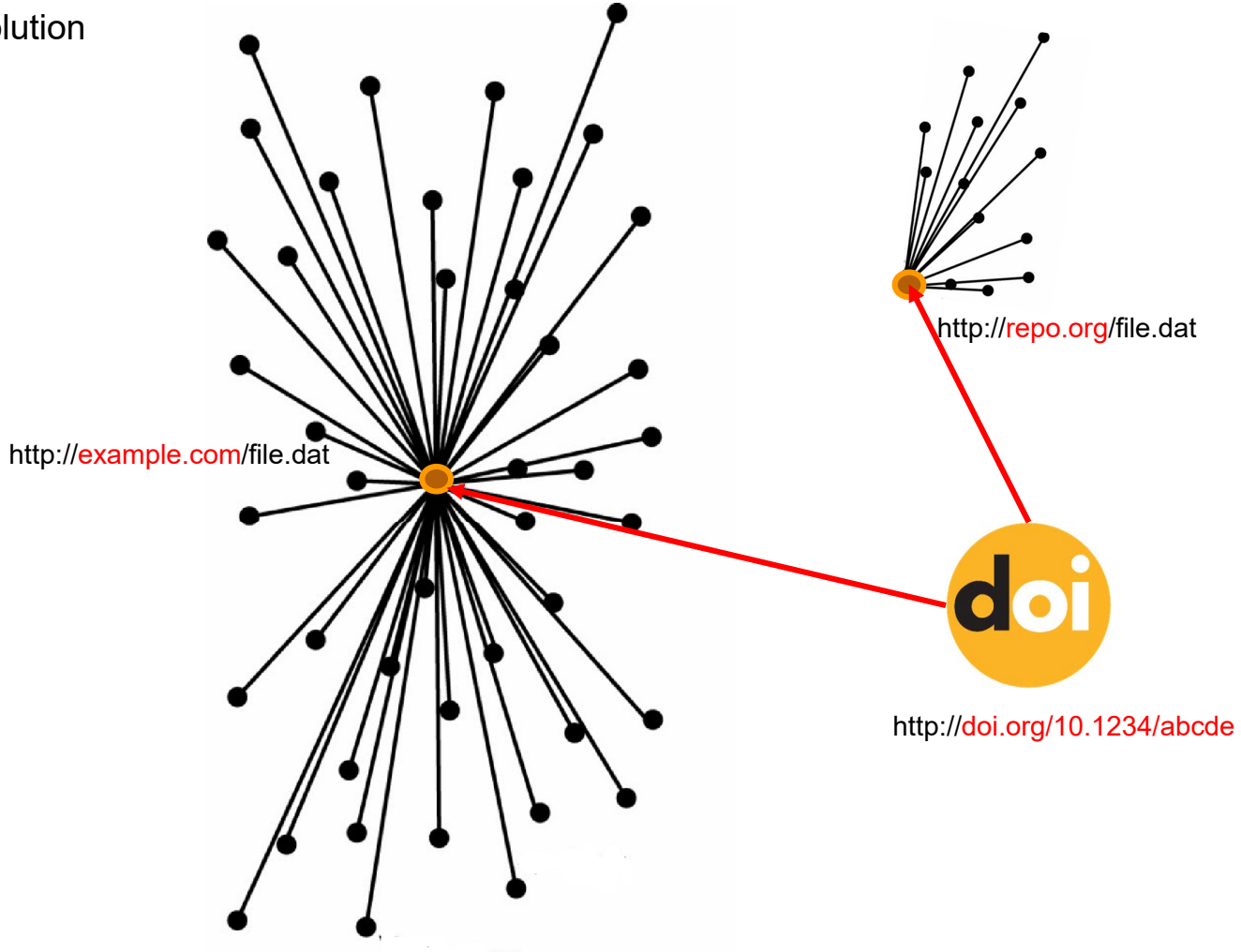

http://example.com/file.dat
file://tmp/file.dat


http://example.com/file.dat
file://tmp/file.dat
file://tmp/file.dat

Web references are mutable & “content negotiable” **by design**



DOI redirection/multi-resolution



Can I find it? Persistent Identifiers and link rot

StackExchange

Where am I supposed to report a broken DOI?

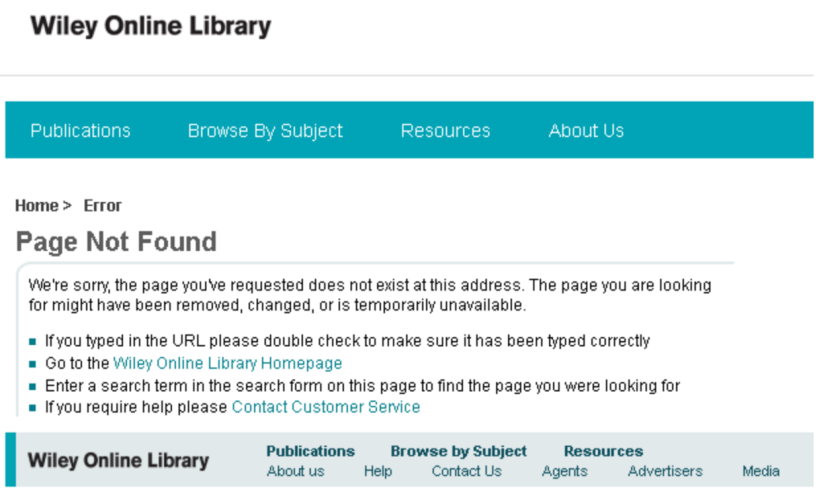
Asked 2 years, 8 months ago Active 2 years, 8 months ago Viewed 688 times

12

Where am I supposed to report a broken DOI? To https://www.doi.org/, to the DOI registration agency that issued the DOI, to whoever is responsible for the website to which the DOI points to, or to somebody else?

For example, http://dx.doi.org/10.1016/0364-0213(90)90002-E is 404:

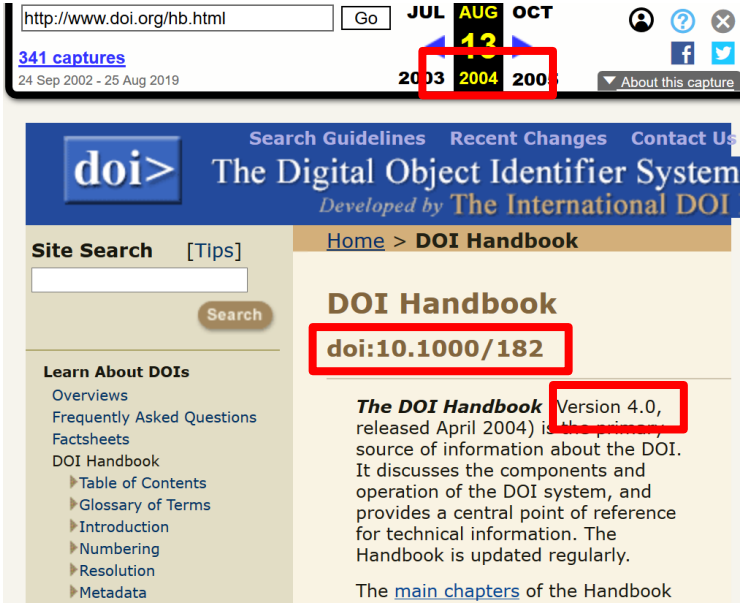
1



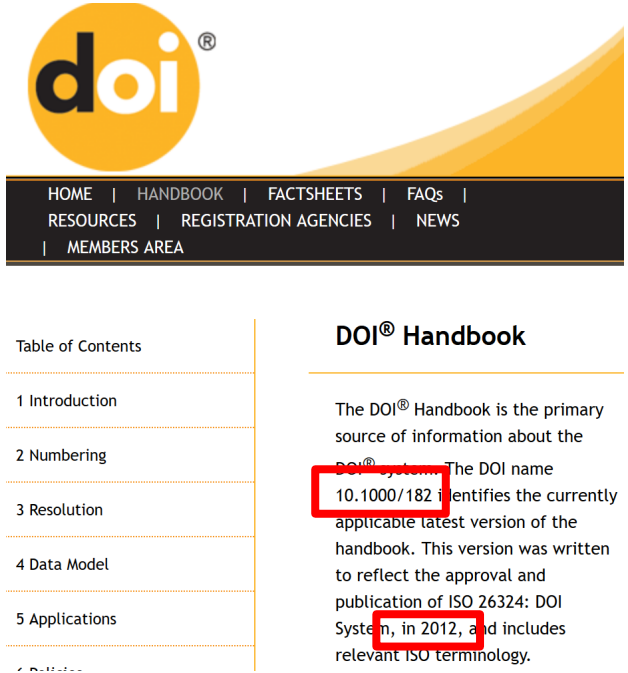
Can I trust it? Persistent Identifiers and content drift

doi: 10.1000/182

2004 (Version 4)



2019 (Version > 6)



Arnap & Hutchinson 2006 10.1145/1179509.1179514

“**persistence is purely a matter of service** and is neither inherent in an object nor conferred on it by a particular naming syntax. The best that an identifier can do is to lead users to the services that support robust reference.”

- J. Kunze et al. “The ARK Identifier Scheme”
https://tools.ietf.org/html/draft-kunze-ark-18

Can I Trust it? Citation and versioning of **Research Data**

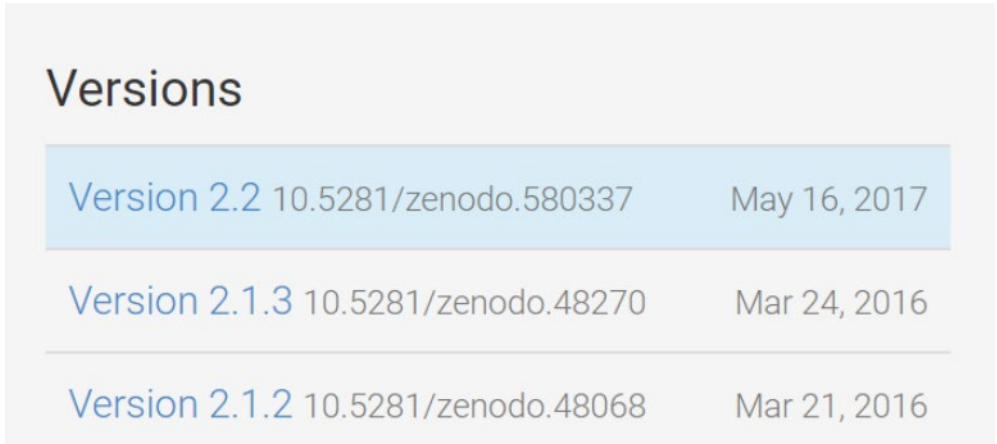
"The demand for **reproducibility** of research results is growing. [There is need] to **reference the exact version of the data** that was used to underpin the research findings, and/or was used to generate higher level products. "

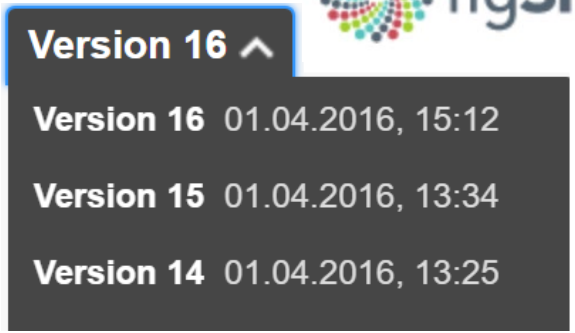
Data Citation WG

Data Versioning WG

Can I trust it? DOI Versioning

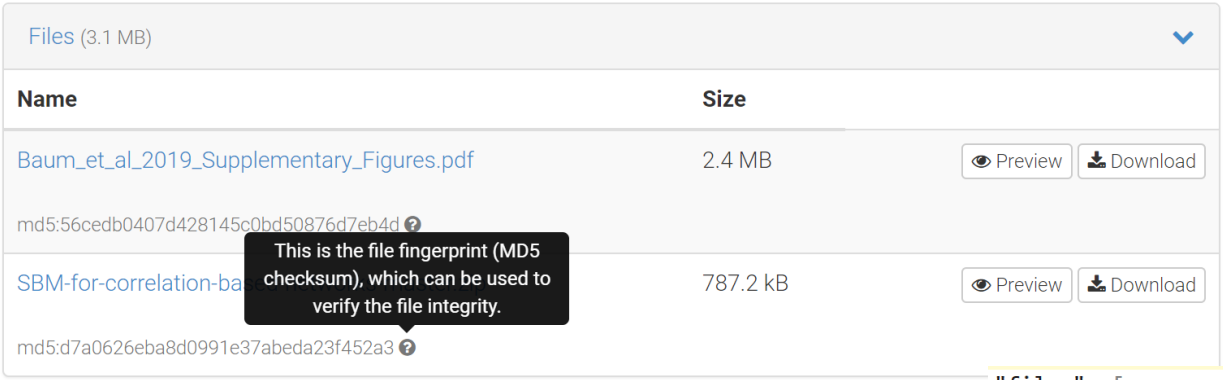
```
{
  "relatedIdentifier": "10.5281/zenodo.580337",
  "relatedIdentifierType": "DOI",
  "relationType": "HasVersion"
}
```





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

Can I trust it? Allowing content verification with checksums



```
"files": [
  {
    "links": {
      "self": "https://zenodo.org/api/files/bd10f6db-e535
        -4436-96f6-d6820009bf85
        /Baum_et_al_2019_Supplementary_Figures.pdf"
    },
    "checksum": "md5:56cedb0407d428145c0bd50876d7eb4d",
    "bucket": "bd10f6db-e535-4436-96f6-d6820009bf85",
    "key": "Baum_et_al_2019_Supplementary_Figures.pdf",
    "type": "pdf",
    "size": 2351627
  },
```

Can I trust it? Content-Addressing for research data?

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

Software Heritage
PRESERVING TECHNICAL KNOWLEDGE

“Identifiers for Digital Objects” rather than “Digital Identifiers of Objects”

Content-addressed PID scheme:
swh:1:cnt:94a9ed024d3859793618152ea559a168bbcbb5e2

Can I trust it? Content-addressed and distributed systems

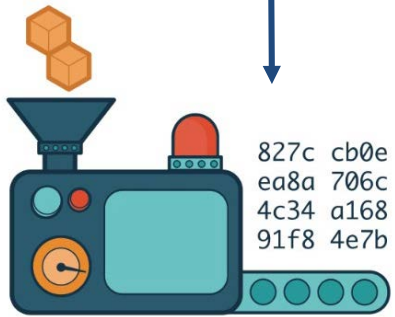
> “it is really about the **ability to trust your data**. I guarantee you, if you put your data in Git, you can trust the fact that five years later after it was converted from your hard disk to your DVD to whatever new technology and you copied it along, **you can verify** that the data you get back out is the exact same as the data you put in.”

“if you cannot guarantee that what I put in […] comes out exactly the same, your system is not worth using”

- Linus Torvalds

LOCATION-ADDRESSING

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

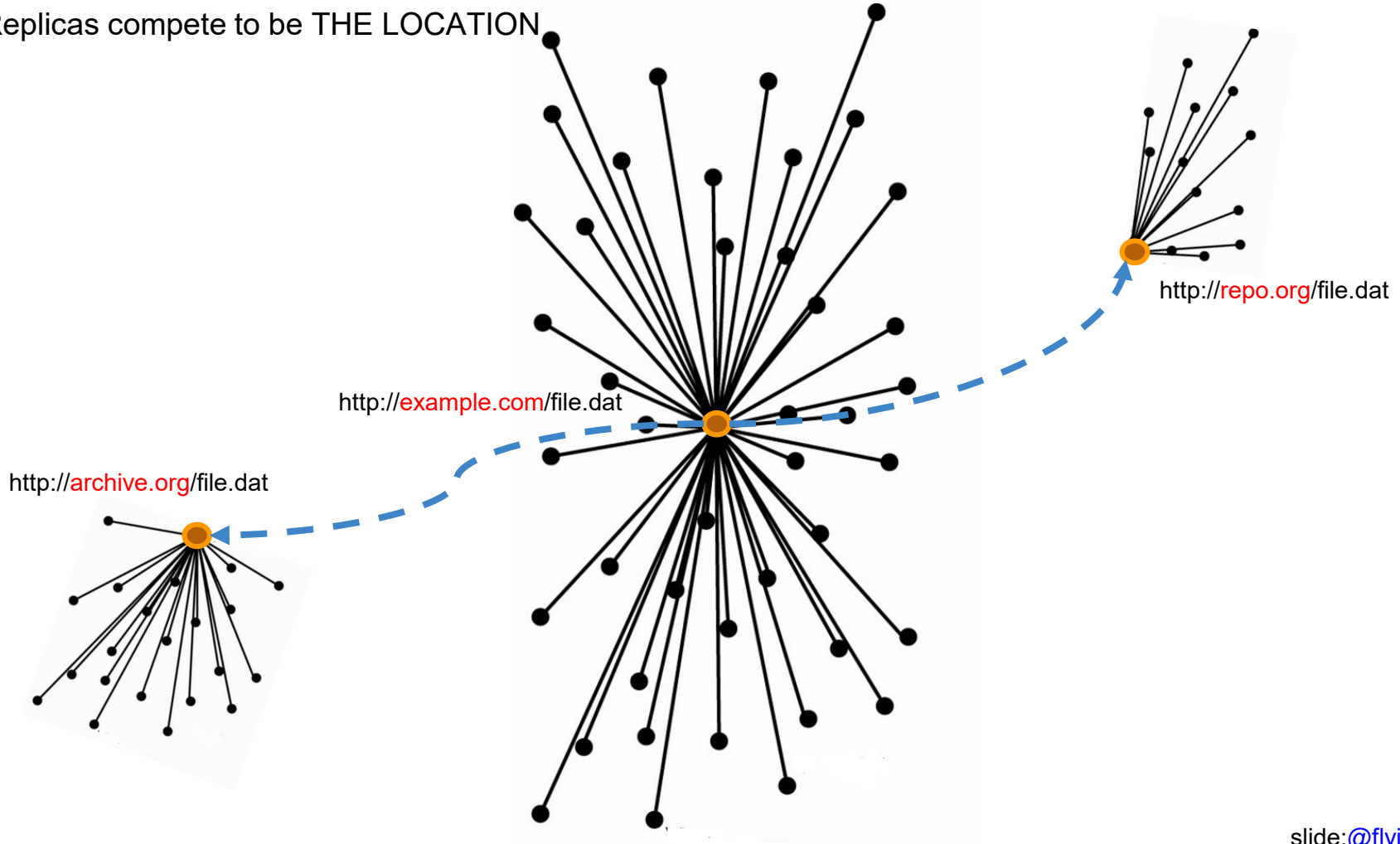


ipfs://zdj7WjqNrjReTcEveRh/pids.pdf

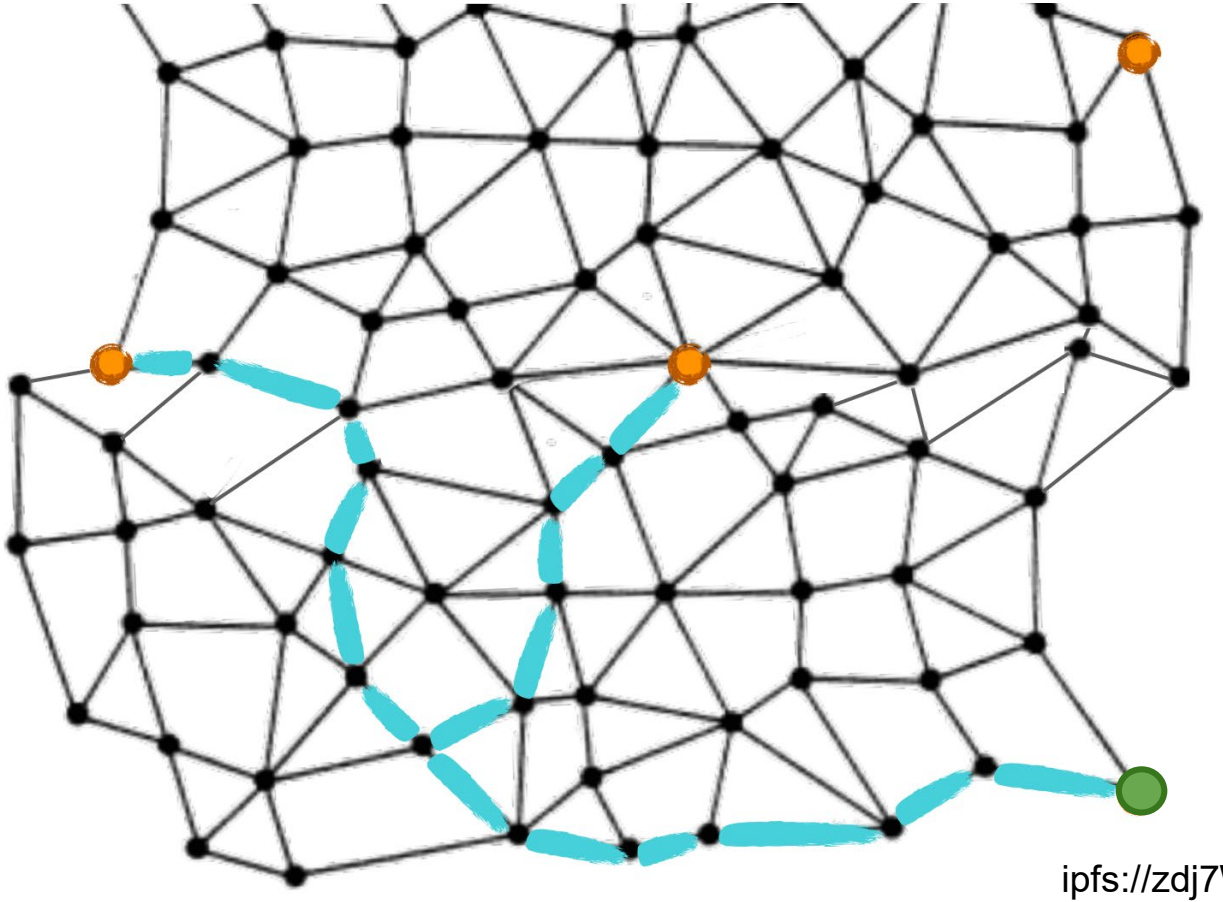
CONTENT-ADDRESSING

Image: @protocollabs

Replicas compete to be THE LOCATION



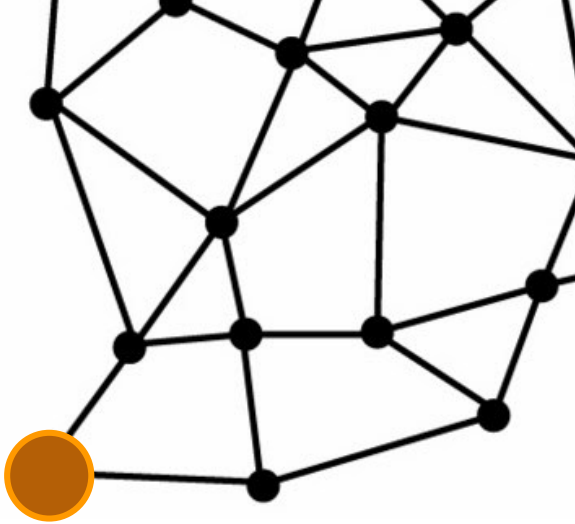
slide:@flyingzumwalt



ipfs://zdj7WjqNrjReTcEveRhgcsXsJvSGwLxJ7js1R7ZCzNaQSKuTh

image:@flyingzumwalt

Great, so let's just use IPFS…?



Immutability != permanent/persistent availability

- Who coordinates 'nodes of last resort' (c.f. LOCKSS, Keepers Registry)?

- Persistent availability of large amounts of research data = **Collective Action Problem** (see: 10.5334/kula.7)

How long is long enough?

- Inevitability of hash collisions (at some stage)?

- For the scholarly record, you still need an indirection layer, to be able to update citations to point at new hashes (sound familiar?)

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

"persistence is purely a matter of service"

Can I trust it? Content-address + Indirection

PID Kernel Information WG

“Kernel record/metadata"

minid

vDOI

Direct access to content associated with a DOI #2

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

Can I trust it? What's the right recipe?


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

title
…
…
contentURL
checksum

"Kernel record/metadata"


```
"files": [
  {
    "links": {
      "self": "https://zenodo.org/api/files/bd10f6db-e535
        -4436-96f6-d6820009bf85
        /Baum_et_al_2019_Supplementary_Figures.pdf"
    },
    "checksum": "md5:56cedb0407d428145c0bd50876d7eb4d",
    "bucket": "bd10f6db-e535-4436-96f6-d6820009bf85",
    "key": "Baum_et_al_2019_Supplementary_Figures.pdf",
    "type": "pdf",
    "size": 2351627
  },
```

Answers?