

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>

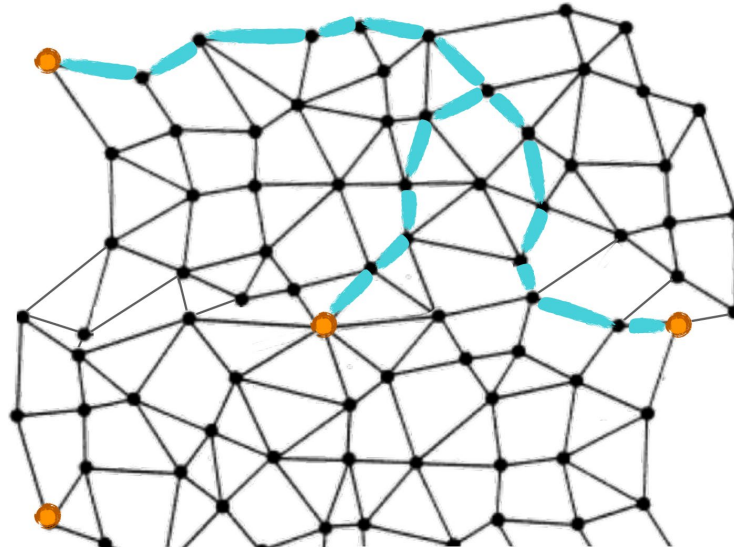
Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

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

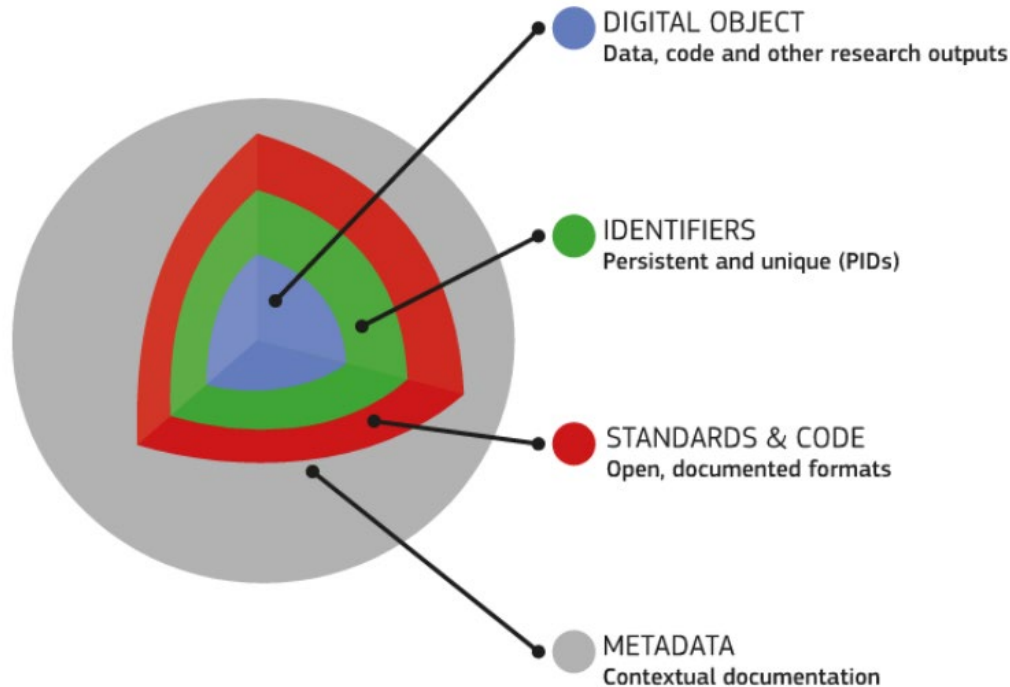


In my 20 years experience in the [#digipres](#) domain I have read a fair share of complex theory, principles, etc.. In practice, I am always able to simplify 'things' and group them under three core questions: 1) can I find it? 2) can I use it? 3) can

I trust it?



Can I Use It? “FAIR Digital Objects”



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

BDBags



researchobject.org

F FRICTIONLESS DATA
STANDARDS AND TOOLING

Psych-DS



Portable
Encapsulated
Project

DataONE
Data Observation Network for Earth

o2r



RO-Crate



DwC Archive

ReproZip

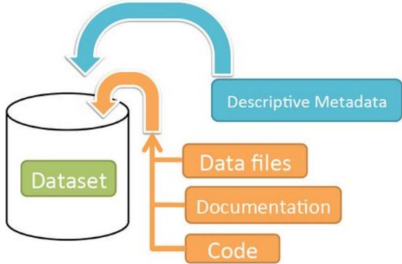


image: @OA_RHUL

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”



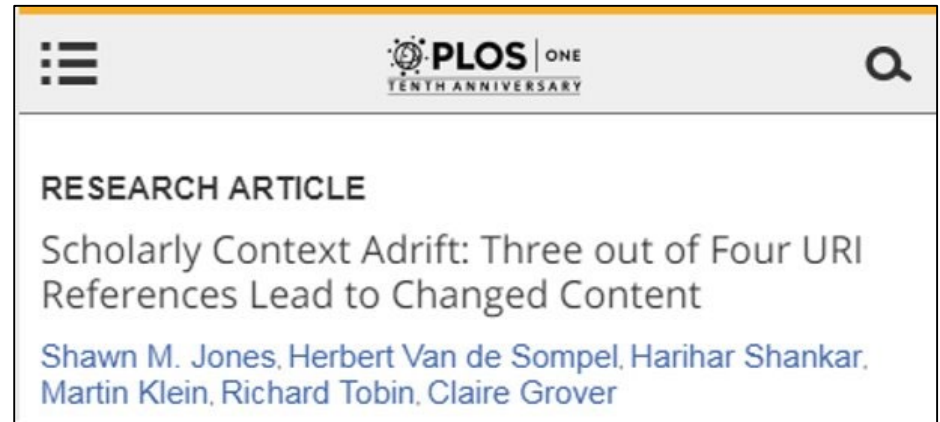
The screenshot shows the top portion of a PLOS ONE article page. At the top left is a hamburger menu icon. In the center is the PLOS ONE logo with 'TENTH ANNIVERSARY' written below it. At the top right is a search icon. Below the header, the text 'RESEARCH ARTICLE' is displayed in bold. The main title is 'Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot'. Below the title, the authors are listed: Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin.

RESEARCH ARTICLE

Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin

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



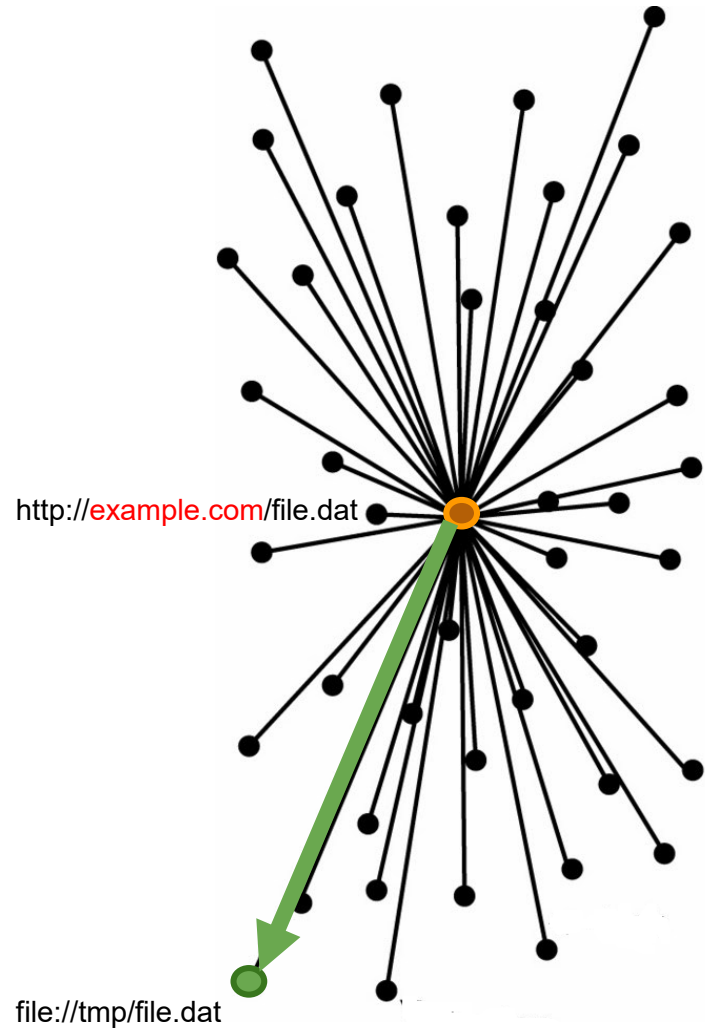
The screenshot shows the top portion of a PLOS ONE article page. At the top left is a hamburger menu icon. In the center is the PLOS ONE logo with 'TENTH ANNIVERSARY' written below it. At the top right is a search icon. Below the header, the text 'RESEARCH ARTICLE' is displayed in bold. The main title is 'Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content'. Below the title, the authors are listed: Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover.

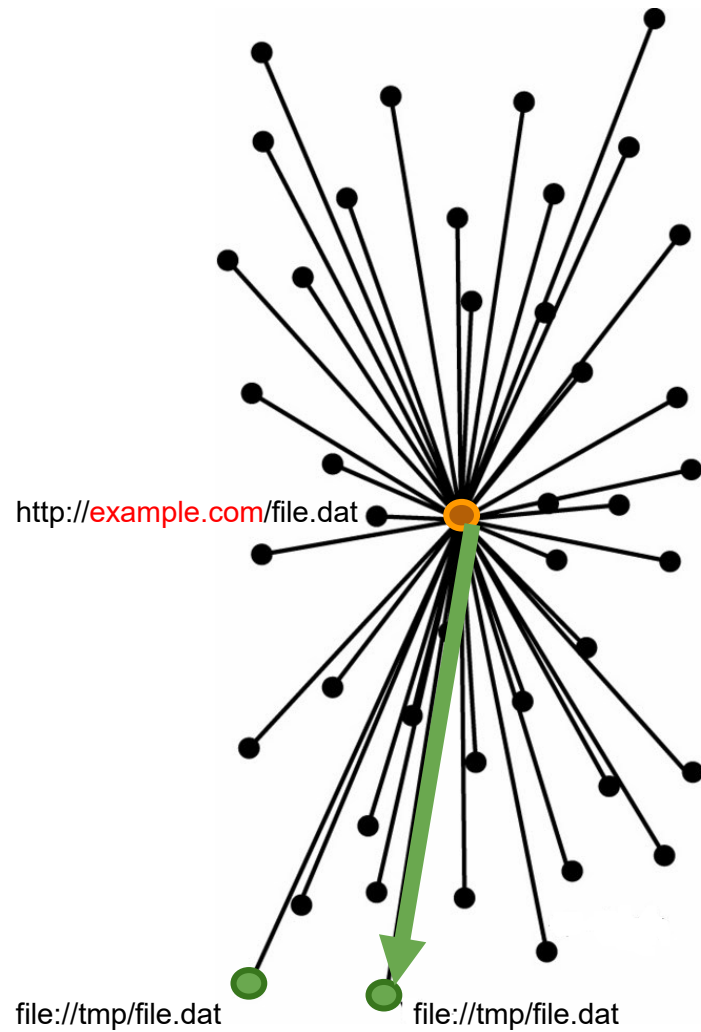
RESEARCH ARTICLE

Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content

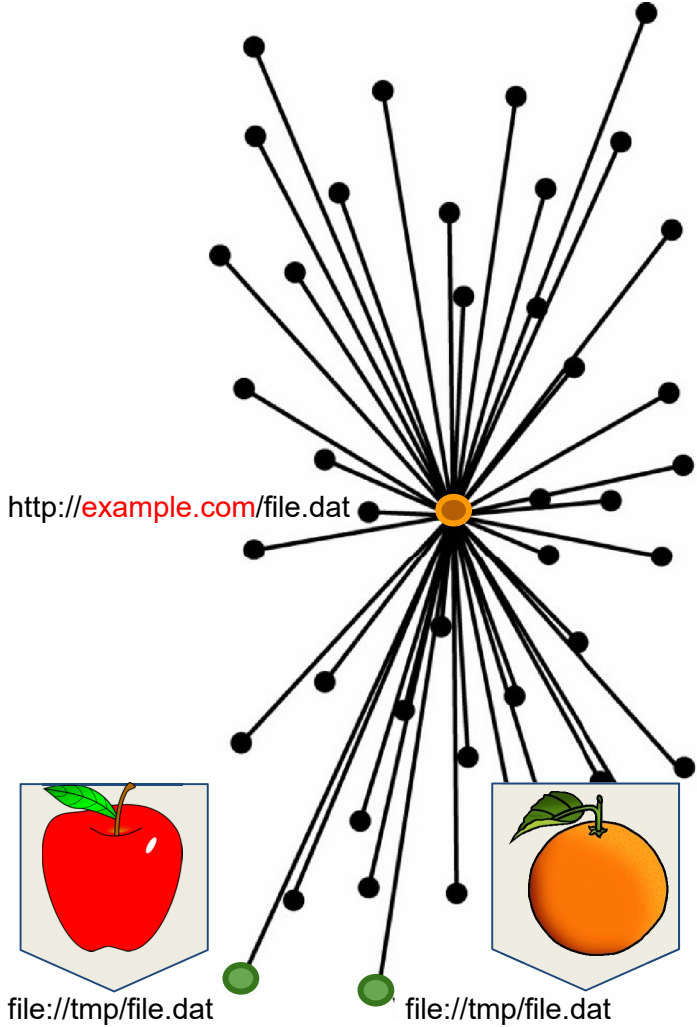
Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, Claire Grover

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



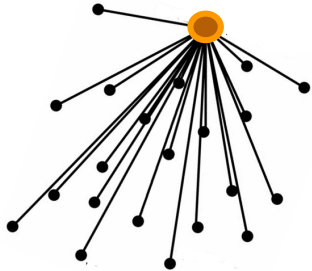


Web references are mutable & “content negotiable” **by design**

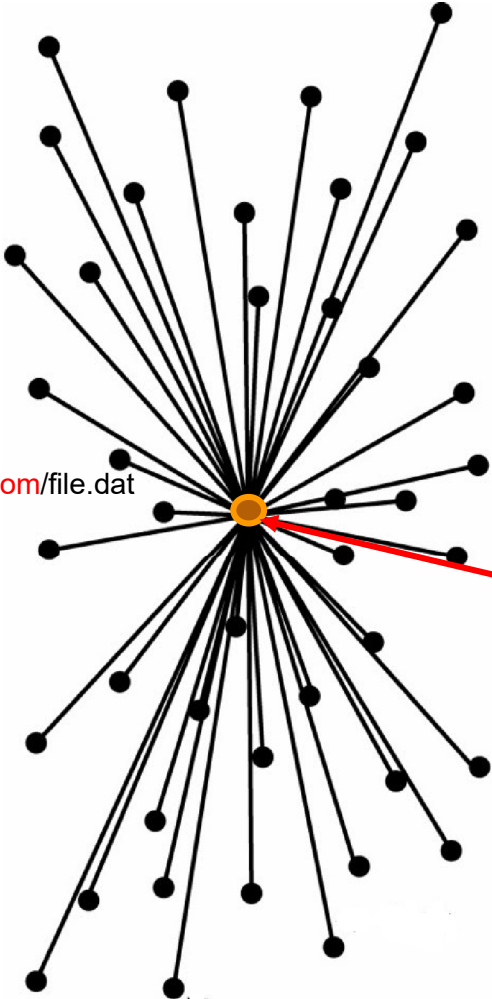


DOI redirection/multi-resolution

<http://archive.net/file.dat>



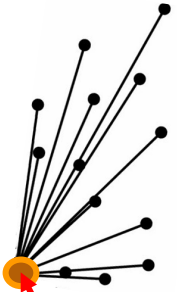
<http://example.com/file.dat>



<http://repo.org/file.dat>



<http://doi.org/10.1234/abcde>



Can I find it? Persistent Identifiers and link rot



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?

▼
★
1

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

Wiley Online Library

Publications

Browse By Subject

Resources

About Us

Home > Error

Page Not Found

We're sorry, the page you've requested does not exist at this address. The page you are looking for might have been removed, changed, or is temporarily unavailable.

- If you typed in the URL please double check to make sure it has been typed correctly
- Go to the [Wiley Online Library Homepage](#)
- Enter a search term in the search form on this page to find the page you were looking for
- If you require help please [Contact Customer Service](#)

Wiley Online Library

Publications

About us

Browse by Subject

Help

Contact Us

Resources

Agents

Advertisers

Media

Can I trust it? Persistent Identifiers and content drift

doi: 10.1000/182

2004 (Version 4)

http://www.doi.org/hb.html Go JUL AUG 13 OCT
341 captures 24 Sep 2002 - 25 Aug 2019 2003 2004 2005 About this capture

Search Guidelines Recent Changes Contact Us
doi> The Digital Object Identifier System
Developed by The International DOI Foundation

Site Search [Tips] Home > DOI Handbook

Search

Learn About DOIs
Overviews
Frequently Asked Questions
Factsheets
DOI Handbook
▶ Table of Contents
▶ Glossary of Terms
▶ Introduction
▶ Numbering
▶ Resolution
▶ Metadata

DOI Handbook
doi:10.1000/182

The DOI Handbook Version 4.0, released April 2004) is the primary source of information about the DOI. It discusses the components and operation of the DOI system, and provides a central point of reference for technical information. The Handbook is updated regularly.

The [main chapters](#) of the Handbook

2019 (Version > 6)

doi®

HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | REGISTRATION AGENCIES | NEWS | MEMBERS AREA

Table of Contents

- 1 Introduction
- 2 Numbering
- 3 Resolution
- 4 Data Model
- 5 Applications

DOI® Handbook

The DOI® Handbook is the primary source of information about the DOI® system. The DOI name **10.1000/182** identifies the currently applicable latest version of the handbook. This version was written to reflect the approval and publication of ISO 26324: DOI System, in 2012, and includes relevant ISO terminology.

“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"  
}
```

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



Version 16 ^

Version 16 01.04.2016, 15:12

Version 15 01.04.2016, 13:34

Version 14 01.04.2016, 13:25

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

Can I trust it? Allowing content verification with checksums



Files (3.1 MB)	
Name	Size
Baum_et_al_2019_Supplementary_Figures.pdf	2.4 MB
md5:56cedb0407d428145c0bd50876d7eb4d ?	
SBM-for-correlation-ba	787.2 kB
md5:d7a0626eba8d0991e37abeda23f452a3 ?	

This is the file fingerprint (MD5 checksum), which can be used to verify the file integrity.

```
"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>

“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



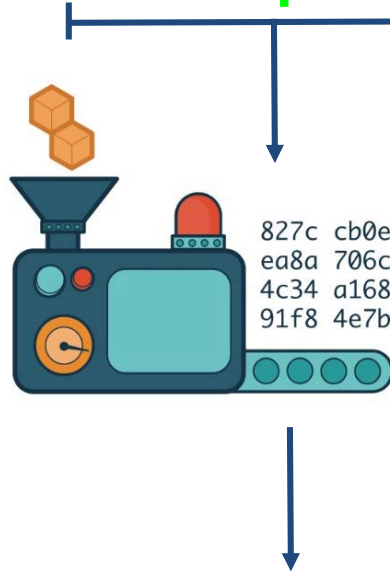
“it is really about the **ability to trust your data**. I guarantee you, if you put your data in Git, you can trust the fact that five years later after it was converted from your hard disk to your DVD to whatever new technology and you copied it along, **you can verify** that the data you get back out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

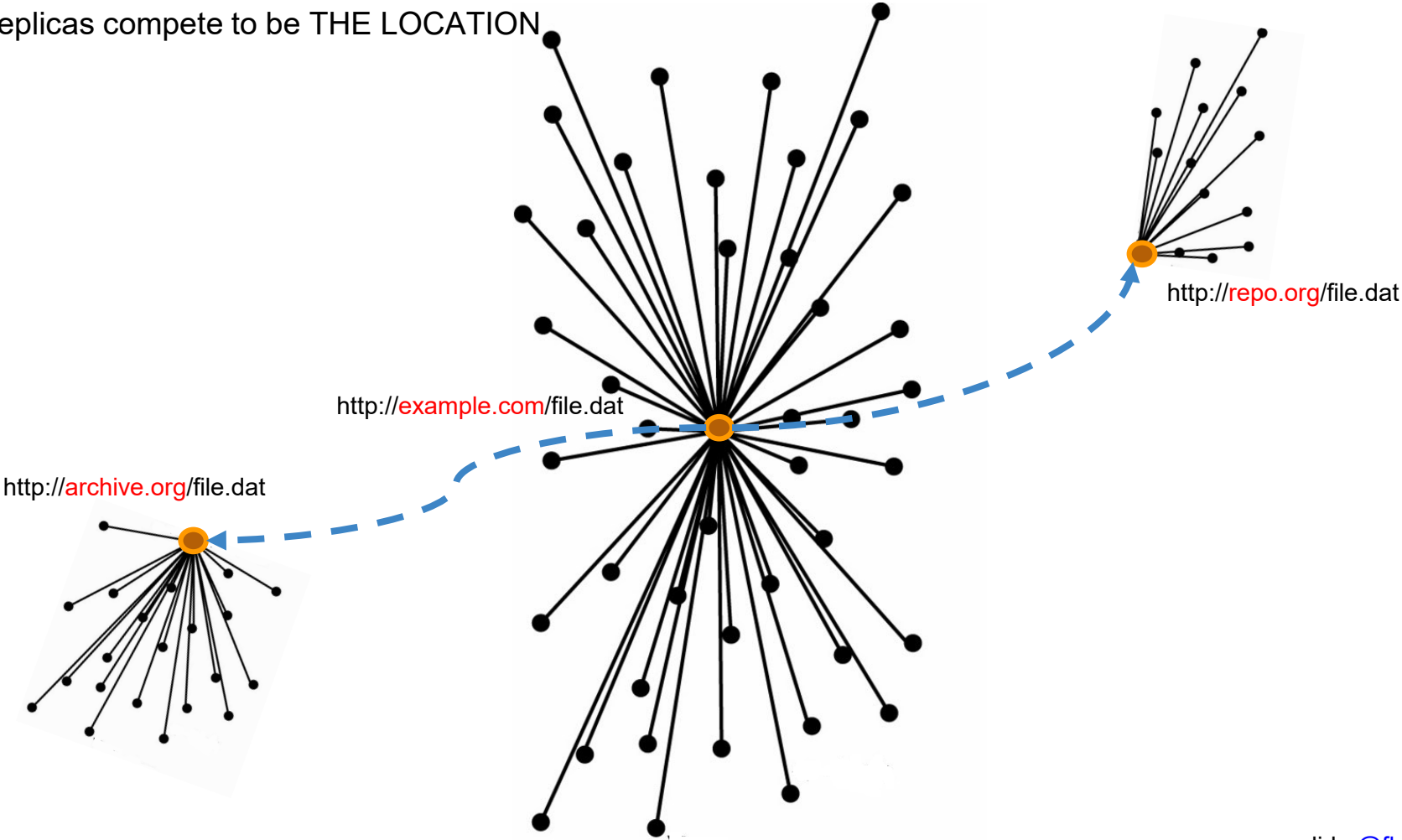
<http://site.com/data/pids.pdf>

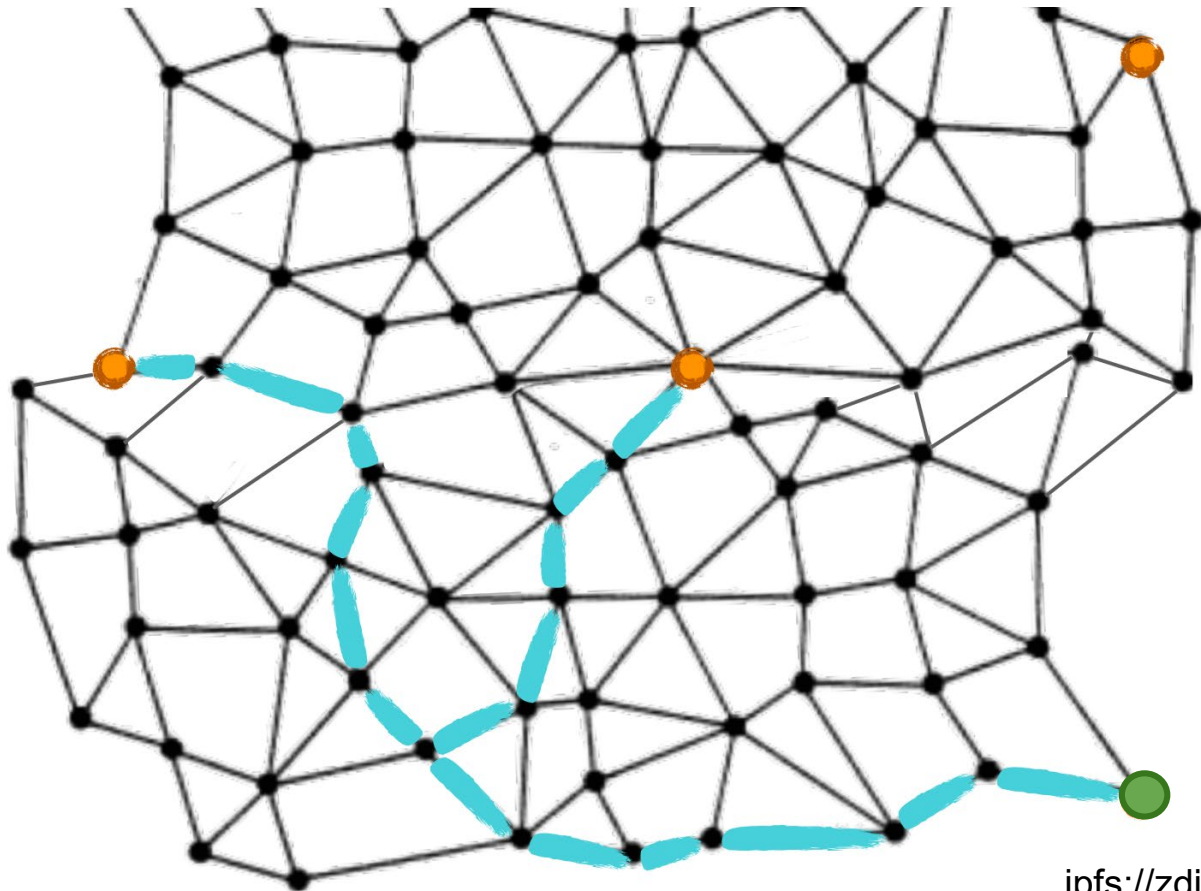


<ipfs://zdj7WjqNryReTcEveRh/pids.pdf>

CONTENT-ADDRESSING

Replicas compete to be THE LOCATION





ipfs://zdj7WjqNnrjReTcEveRh
gcsXsJvSGwLxJ7js1R7ZCz
NaQSKuTh

Great, so let's just use IPFS...?

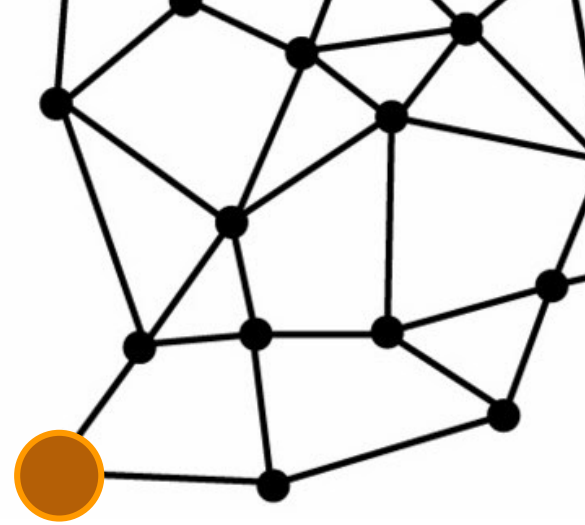
Immutability != permanent/persistent availability

- 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”

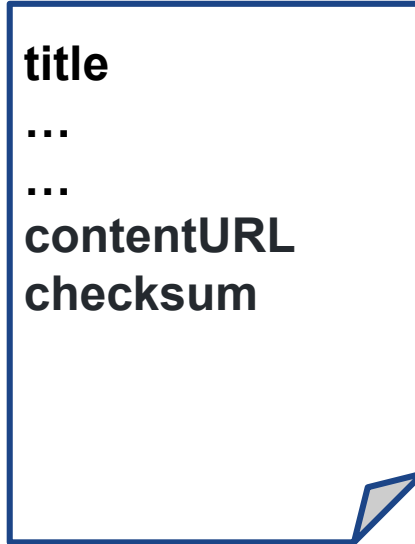


Can I trust it? Content-address + Indirection



PID Kernel Information WG

vDOI



“Kernel record/metadata”

minid



Direct access to content associated with a DOI #2



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"  
}
```



“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?