A Decentralized Network for Publishing Linked Data

—

Nanopublications, Trusty URIs, and Science Bots

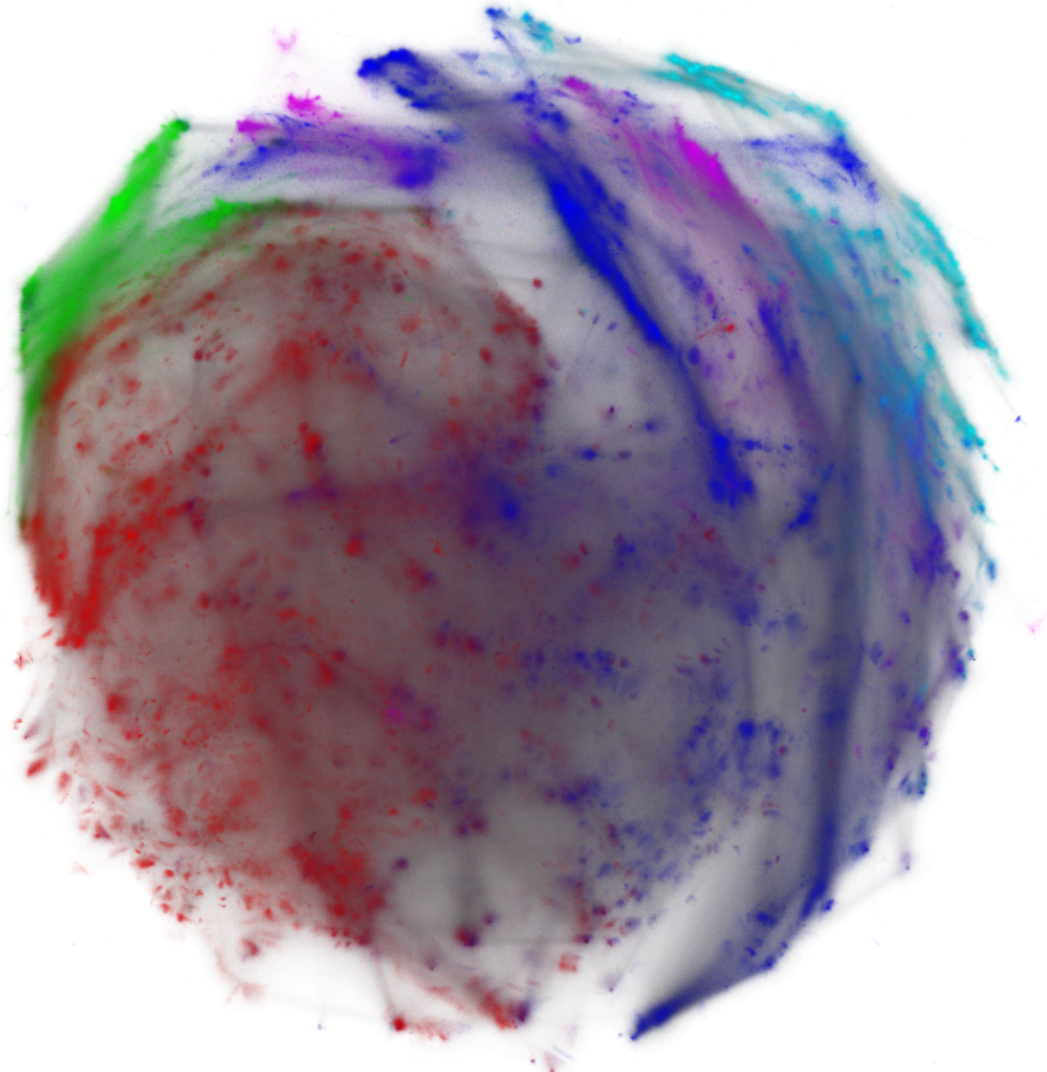
Tobias Kuhn

http://www.tkuhn.ch
@txkuhn

ETH Zurich

CERN Workshop on Innovations in Scholarly Communication
(OAI9)
Geneva
17 June 2015

Increasing Scientific Output: >1.5M New Articles Per Year



Citation network of 30M scientific publications

Image from: Kuhn et al. Inheritance patterns in citation networks reveal scientific memes. *Physical Review X* 4. 2014.

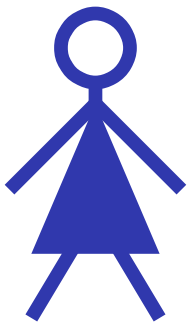
Increasing Importance of Scientific Data

London Underground staff sorting 4M used tickets to analyse line use in 1939

Image from: http://www.telegraph.co.uk/travel/picturegalleries/9791007/The-history-of-the-Tube-in-pictures-150-years-of-London-Underground.html?frame=2447159

Problem:
Replication and Re-Use of Research Results

Exemplary Situation: Sue publishes a script that should allow everybody to replicate her scientific analysis:



```
# Download data:
wget http://some-third-party.org/dataset/1.4
# Analyze data:
...
```

Problems:

- What if the resource becomes unavailable at this location?
- What if the third party silently changes that version of the dataset?
- What if the web site gets hacked and the data manipulated?

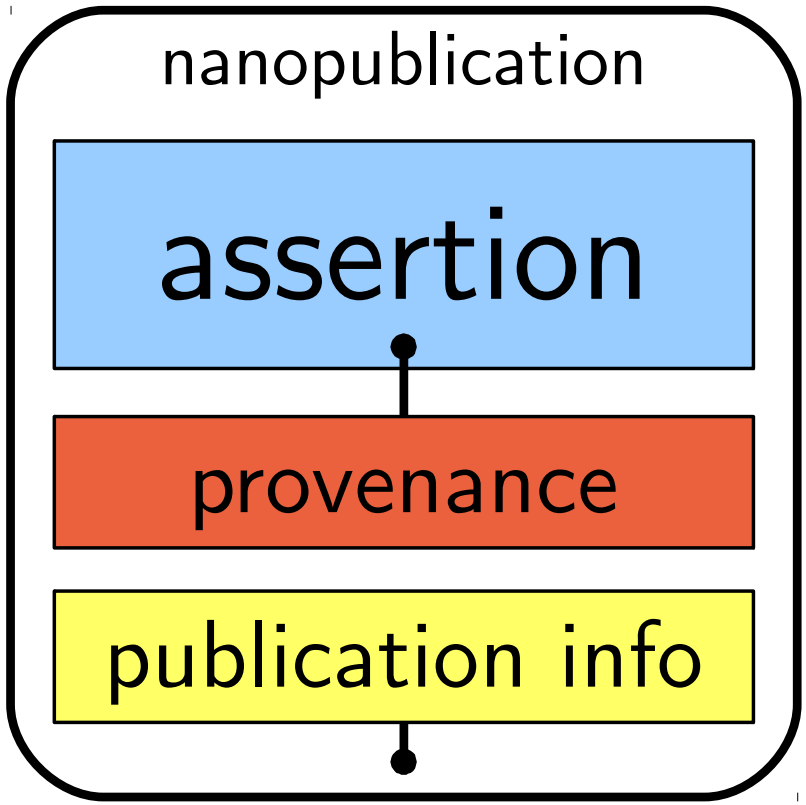
Data Publishing, Archiving, and Re-Use

Scientific datasets become increasingly important, and these data are increasingly produced and consumed directly by software.

Published data should therefore be:

- Verifiable (Is this really the data I am looking for?)
- Immutable (Can I be sure that it hasn't been modified?)
- Permanent (Will it be available in 1, 5, 20 years from now?)
- Reliable (Can it be efficiently retrieved whenever needed?)
- Granular (Can I refer to individual data entries?)
- Semantic (Can it be automatically interpreted?)
- Linked (Does it use established identifiers and ontologies?)
- Trustworthy (Can I trust the source?)

Nanopublications: Provenance-Aware Semantic Publishing (based on RDF)



`http://nanopub.org` / `@nanopub_org`

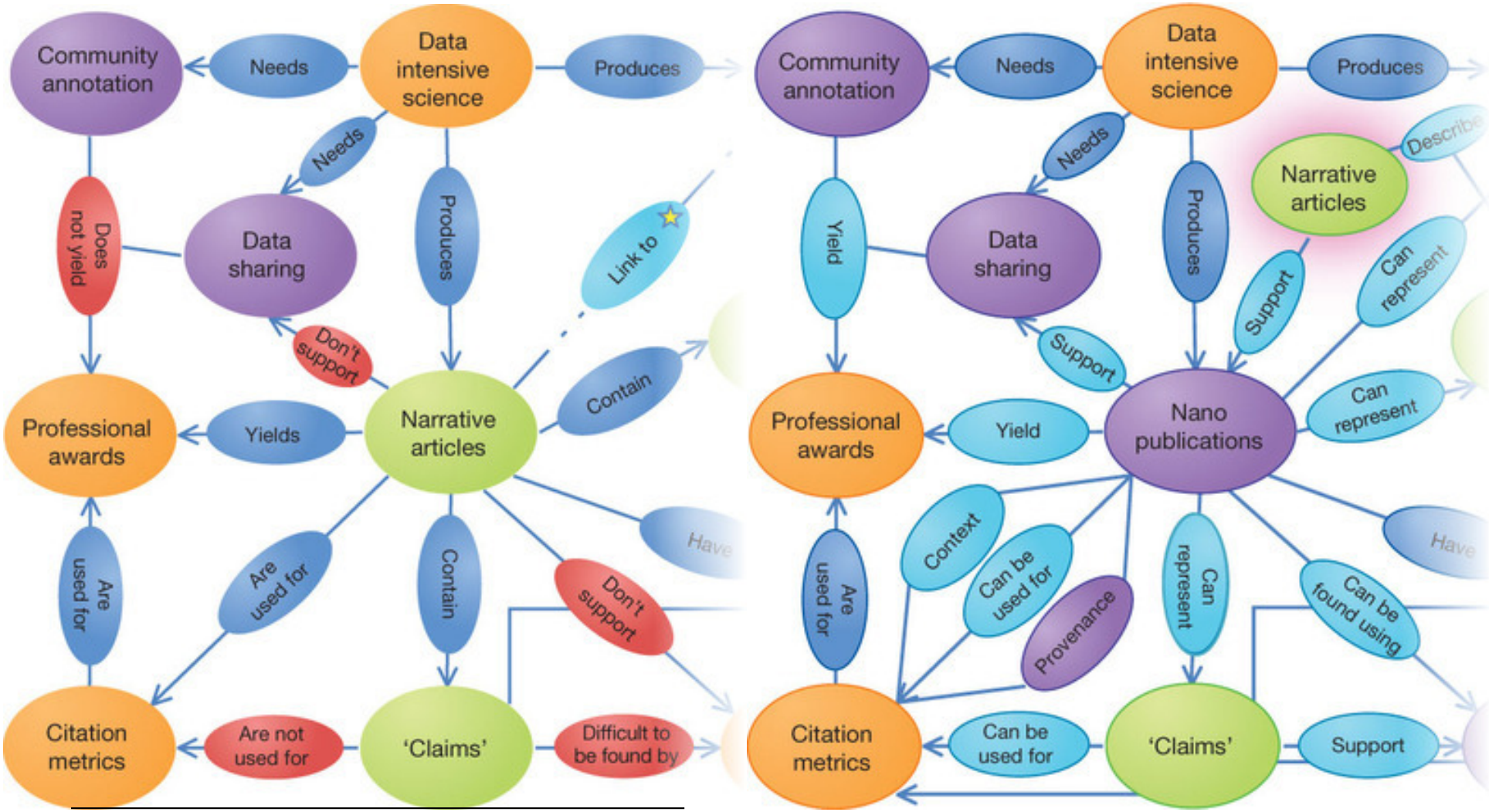
Vision: Changing Scholarly Communication

Now

Narrative articles at the center

Future

Nanopublications at the center



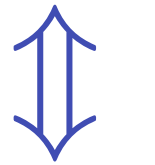
Images from Mons et al. The value of data. *Nature genetics*, 43(4):281–283, 2011

Nanopublication Example

```
sub:assertion {
  sub:_3 a rdf:Statement ; rdf:subject schem:Adenosine%20triphosphate ;
    rdf:predicate belv:decreases ; rdf:object sub:_1 ;
    occursIn: obo:UBERON_0001134 , species:9606 .
  sub:_1 a go:0003824 ; hasAgent: sub:_2 .
  sub:_2 a Protein: ; geneProductOf: hgnc:12517 .
}
sub:provenance {
  sub:assertion prov:hadPrimarySource pubmed:9703368 ;
    prov:wasDerivedFrom beldoc: , sub:_4 .
  beldoc: dce:description "Approximately 61,000 statements." ;
    dce:rights "Copyright (c) 2011-2012, Selventa. All rights reserved." ;
    dce:title "BEL Framework Large Corpus Document" ;
    pav:authoredBy sub:_5 ; pav:version "20131211" .
  sub:_4 prov:value "UCP1 contains six potential transmembrane a-helices (72) an
    prov:wasQuotedFrom pubmed:9703368 .
  sub:_5 rdfs:label "Selventa" .
}
sub:pubinfo {
  this: dct:created "2014-07-03T14:34:13.226+02:00"^^xsd:dateTime ;
    pav:createdBy orcid:0000-0001-6818-334X , orcid:0000-0002-1267-0234 .
}
```

Identifiability Problem of URIs (Web Links)

`http://some-third-party.org/dataset/1.4`



?

Given a URI for a digital artifact, there is no reliable standard procedure of checking whether a retrieved file really represents the correct and original state of that artifact.

Solution: Identifiers that include (iterative) cryptographic hash values (as applied, for example, by Git and Bitcoin)

Cryptographic Hash Values

A cryptographic hash value is a short random-looking sequence of bytes calculated on a given input:

`This is some text.` $\Rightarrow$ `hRUvOM`

The same input always leads to exactly the same value:

`This is some text.` $\Rightarrow$ `hRUvOM`

Even a minimally modified input leads to a completely different value:

`This is xome text.` $\Rightarrow$ `sCtYbf`

The input is not reconstructible from the hash value (in practice):

`This is some text.` $\not\Leftarrow$ `hRUvOM`

Given an input and a matching hash value, we can therefore be perfectly sure that it was exactly that input that led to the hash.

Iterative Hashing

Hash values can be used as identifiers in an iterative fashion:

`This is some text.`	$\Rightarrow$	`hRUvOM`
`This text is based on hRUvOM.`	$\Rightarrow$	`LwGqwX`
`This depends on LwGqwX.`	$\Rightarrow$	`civRbq`

From a single identifier (such as `civRbq`), the entire reference tree can be verified:

`This is some text.`	$\Rightarrow$	✓	`hRUvOM`
`This text is based on hRUvOM.`	$\Rightarrow$	✓	`LwGqwX`
`This depends on LwGqwX.`	$\Rightarrow$	✓	`civRbq`

And any modification can be noticed:

`This is xome text.`	$\Rightarrow$	✗	`hRUvOM`
`This text is based on hRUvOM.`	$\Rightarrow$	✓	`LwGqwX`
`This depends on LwGqwX.`	$\Rightarrow$	✓	`civRbq`

Trusty URIs: Cryptographic Hash Values for Verifiable and Immutable Web Identifiers

Basic idea: Use of cryptographic hash values together with URIs as identifiers for digital artifacts such as nanopublications.

Requirements:

- To allow for the verification of entire reference trees, the hash should be part of the reference (i.e. the URI)
- To allow for meta-data, digital artifacts should be allowed to contain self-references (i.e. their own URI)
- Format-independent hash for different kinds of content (e.g. RDF)
- The complete approach should be decentralized and open
- We want to use them right away

Kuhn, Dumontier. Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data. ESWC 2014.

Verifiable — Immutable — Permanent

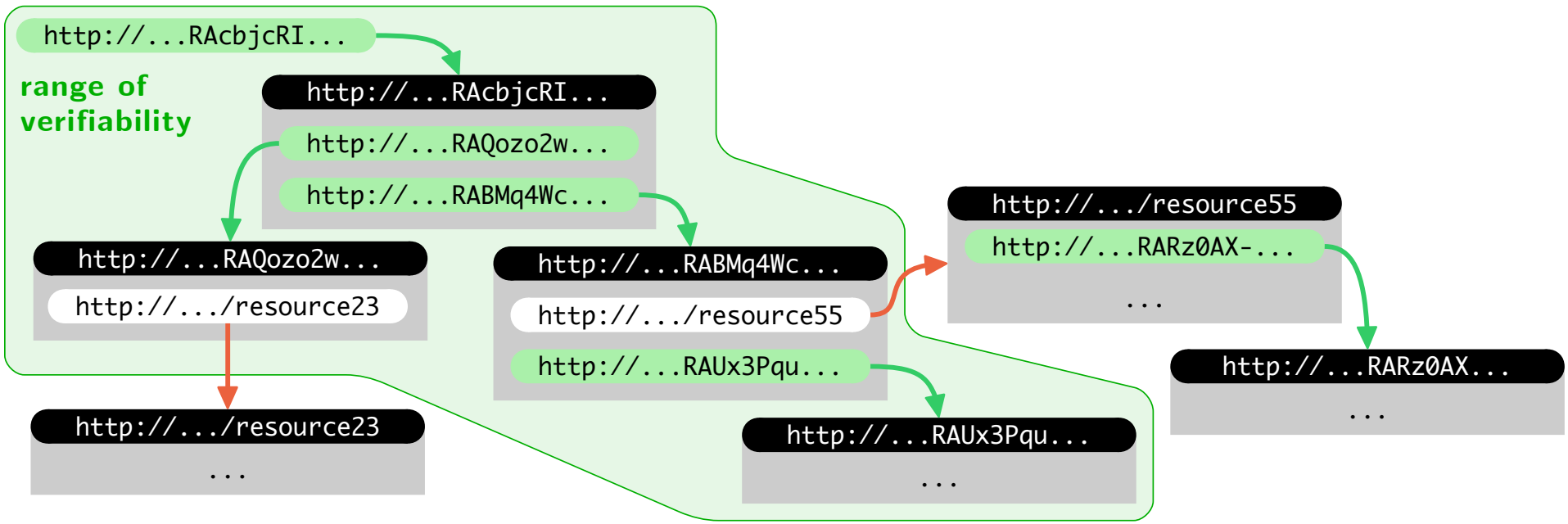
Whether or not a given resource is the one a given trusty URI is supposed to represent can be **verified with perfect confidence**.

(assuming that the trusty URI for the required artifact is known, e.g. because another artifact contains it as a link)

`http://trustyuri.net`

Kuhn, Dumontier. Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data. ESWC 2014.

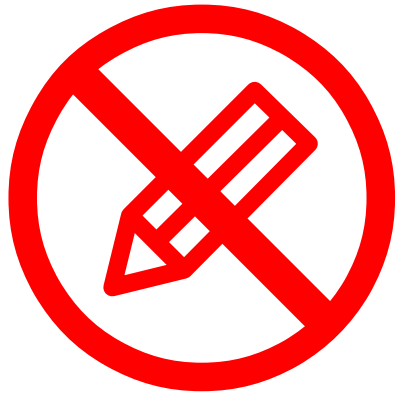
Extended Range of Verifiability Through Iterative Hashing



http://trustyuri.net

Kuhn, Dumontier. Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data. ESWC 2014.

Verifiable — **Immutable** — Permanent



Trusty URI artifacts are **immutable**, as any change in the content also changes its URI, thereby making it a new artifact.

(as soon as your trusty URI has been picked up by third parties, e.g. cached or linked from other resources, every change will be noticed)

`http://trustyuri.net`

Kuhn, Dumontier. Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data. ESWC 2014.

Verifiable — Immutable — **Permanent**

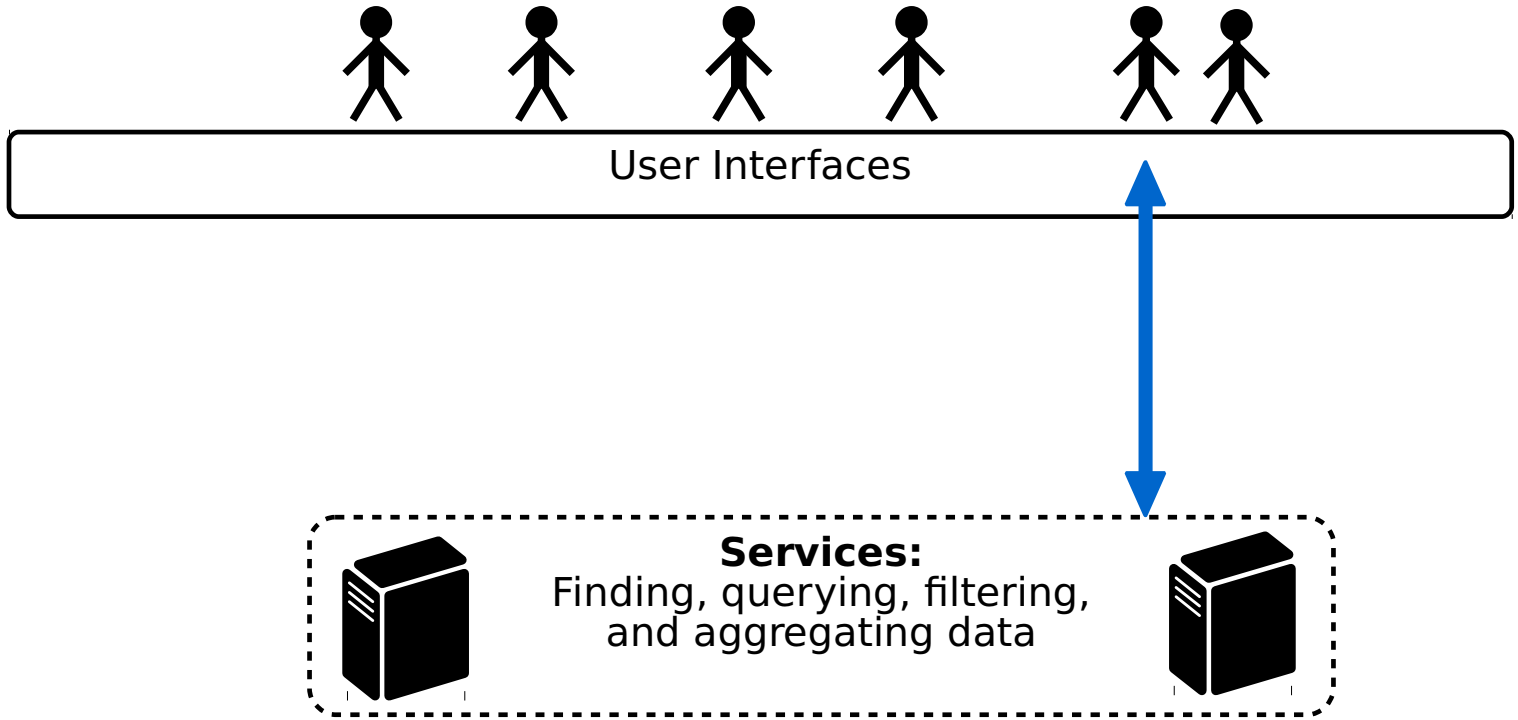
Trusty URI artifacts are **permanent**, as they can be retrieved from the cache of third-party websites if otherwise no longer available.

(if there are services regularly crawling and caching the artifacts on the web)

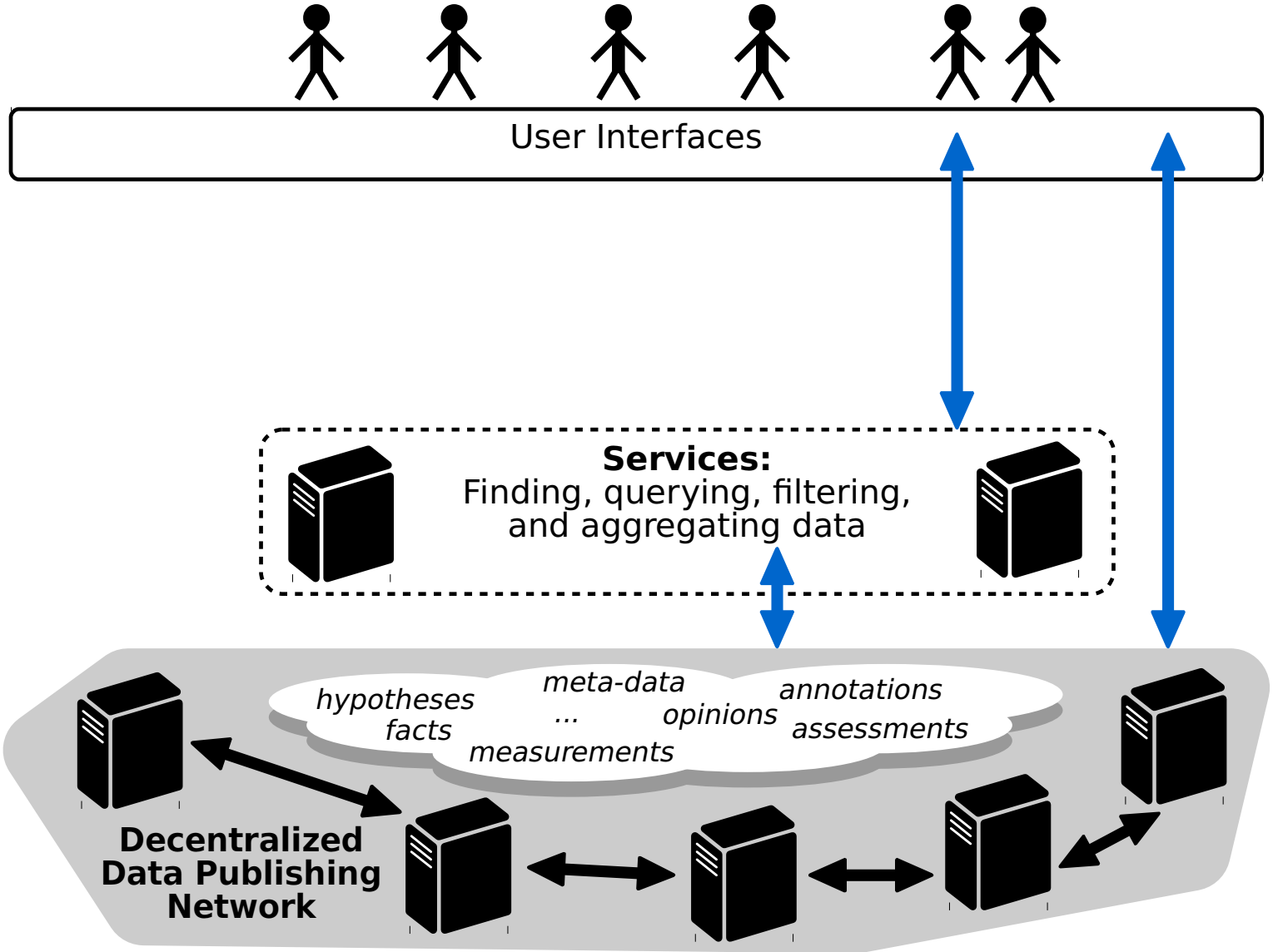
`http://trustyuri.net`

Kuhn, Dumontier. Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data. ESWC 2014.

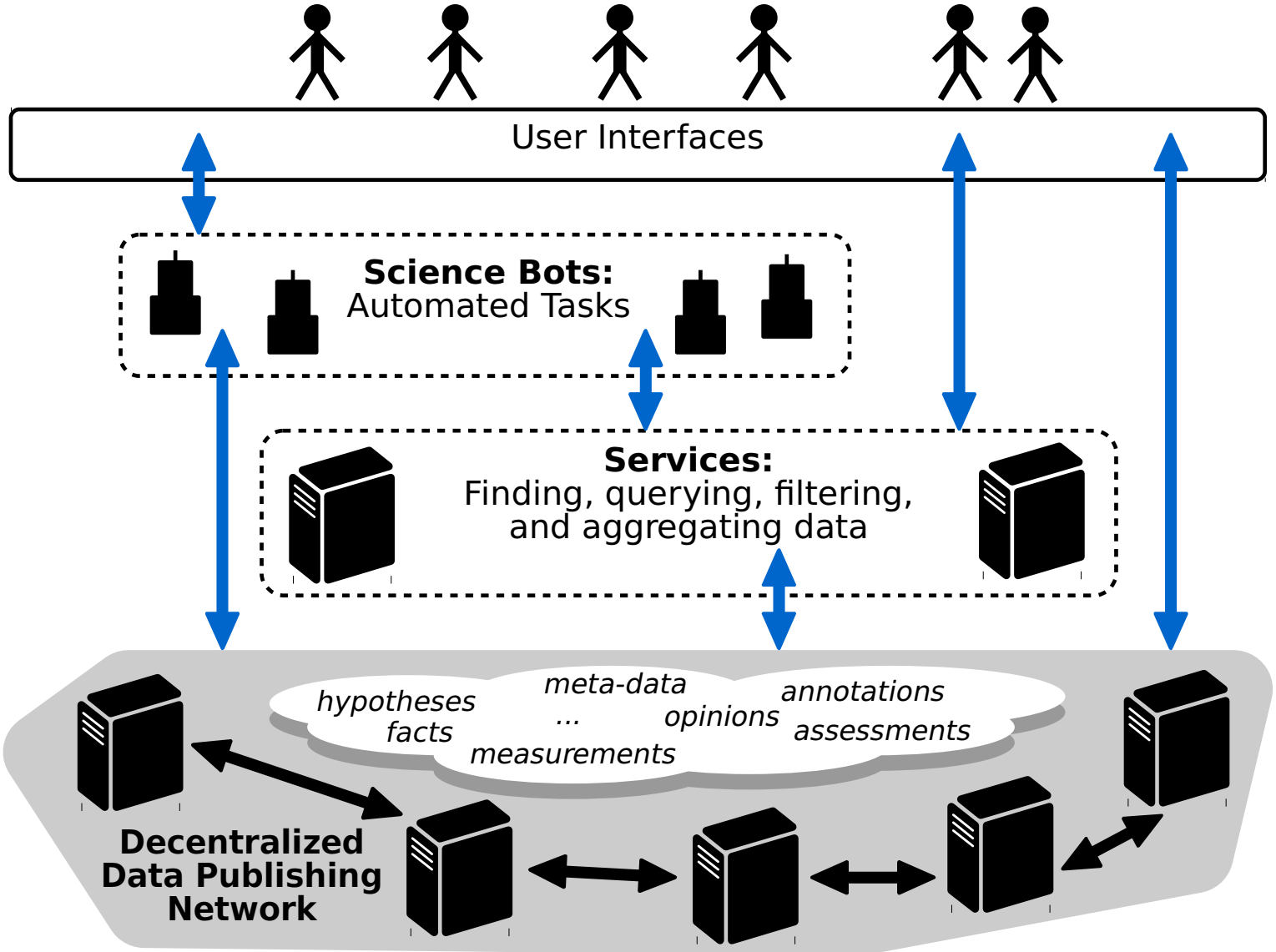
A Multi-Layer Architecture for Reliable Scientific Data Publishing?



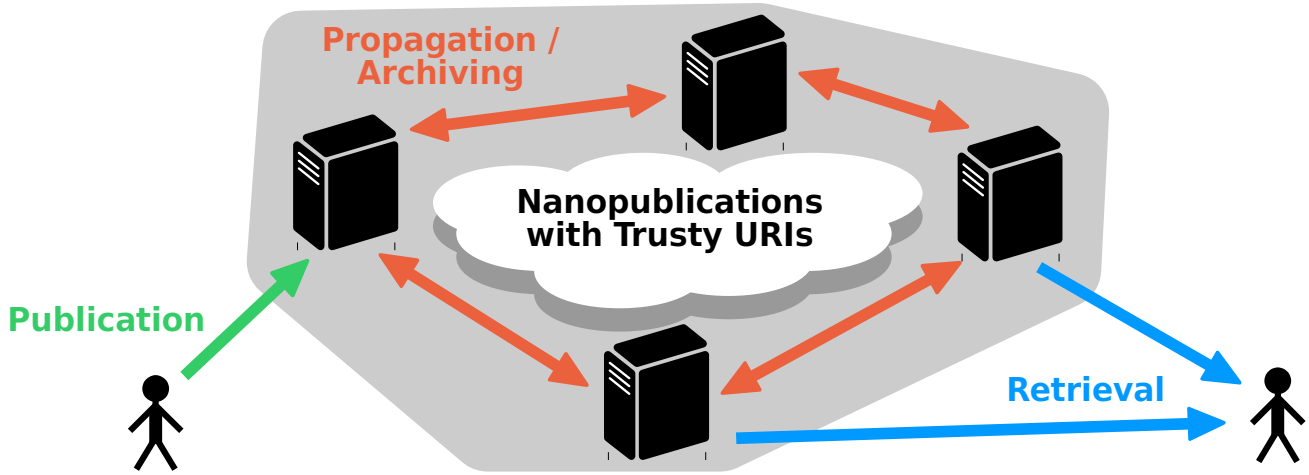
A Multi-Layer Architecture for Reliable Scientific Data Publishing?



A Multi-Layer Architecture for Reliable Scientific Data Publishing?



Decentralized and Reliable Publishing with a Nanopublication Server Network



http://npmonitor.inn.ac

Kuhn et al. Publishing without Publishers: a Decentralized Approach to Dissemination, Retrieval, and Archiving of Data. arXiv:1411.2749.

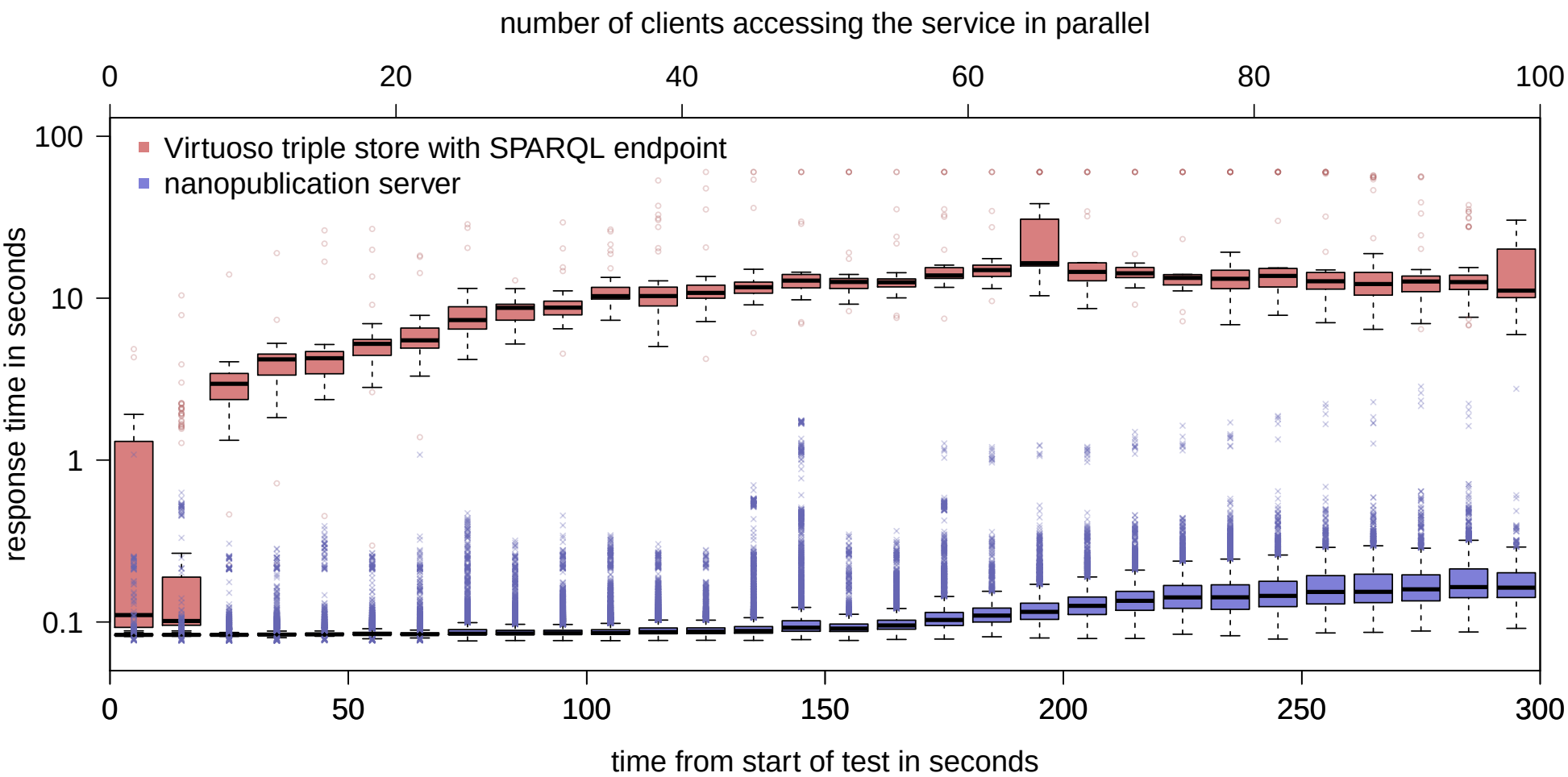
Decentralized — Open — Real-time

- No a central authority: Everybody can set up a server and join the network
- No restrictions on publication: Everybody can upload nanopublications
- No delay between submission and publication: Nanopublications are made public immediately

- No updates: If a nanopublication is modified, that makes it a *new* nanopublication (enforced by trusty URIs)
- No queries: Only simple identifier-based lookup

Kuhn et al. Publishing without Publishers: a Decentralized Approach to Dissemination, Retrieval, and Archiving of Data. arXiv:1411.2749.

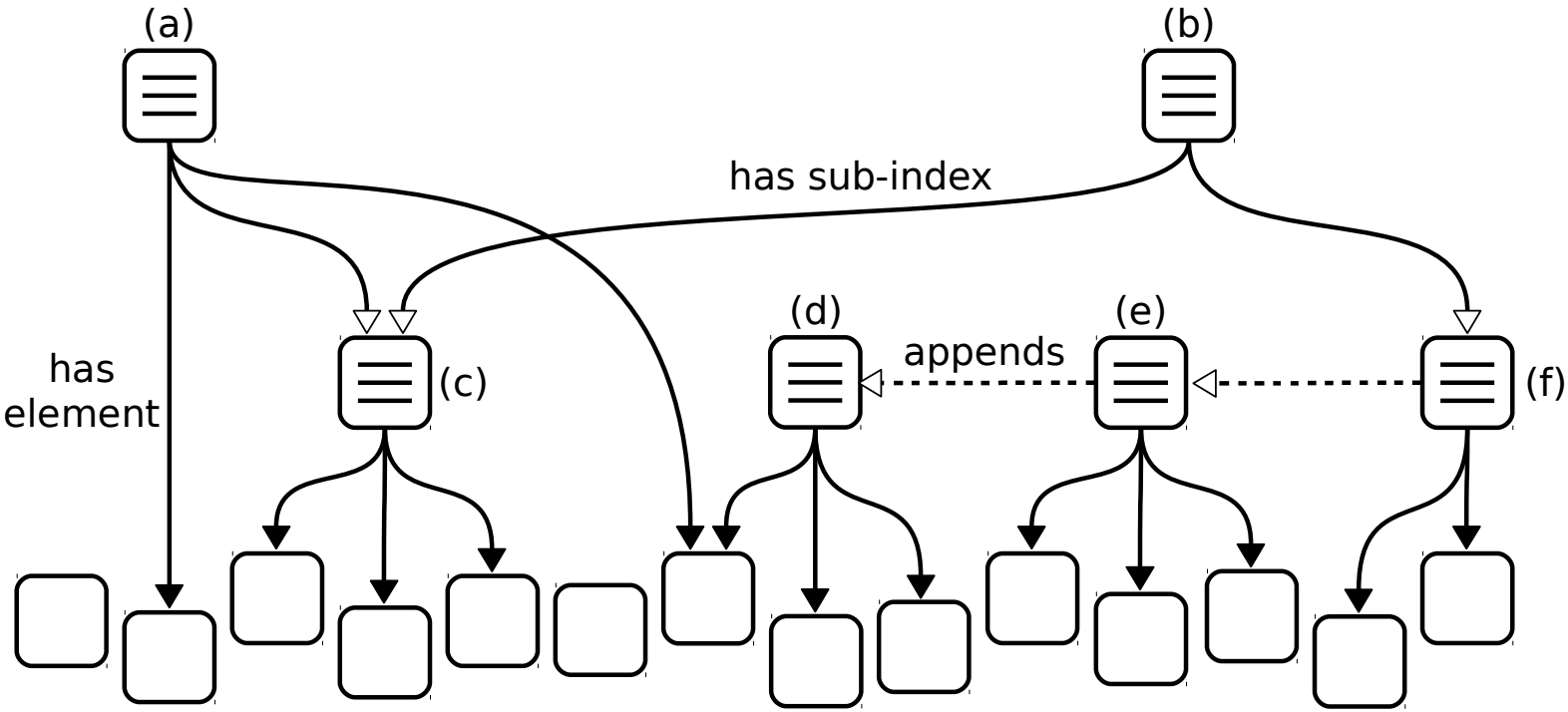
Fast Parallel Access



Kuhn et al. Publishing without Publishers: a Decentralized Approach to Dissemination, Retrieval, and Archiving of Data. arXiv:1411.2749.

Defining Datasets with Nanopublication Indexes

(which are themselves Nanopublications)



Kuhn et al. Publishing without Publishers: a Decentralized Approach to Dissemination, Retrieval, and Archiving of Data. arXiv:1411.2749.

Using Nanopublication Datasets

Once published in the network, nanopublication indexes can be cited:

[7] Nanopubs converted from OpenBEL's Small and Large Corpus 20131211. Nanopublication index, 4 March 2014, `http://np.inn.ac/RAR5dwELYLKGSfrOclnWhjOj-2nGZN_8BW1JjxwFZINHw`

Researchers can then fetch and reuse the data in a reliable and prefectly reproducible manner:

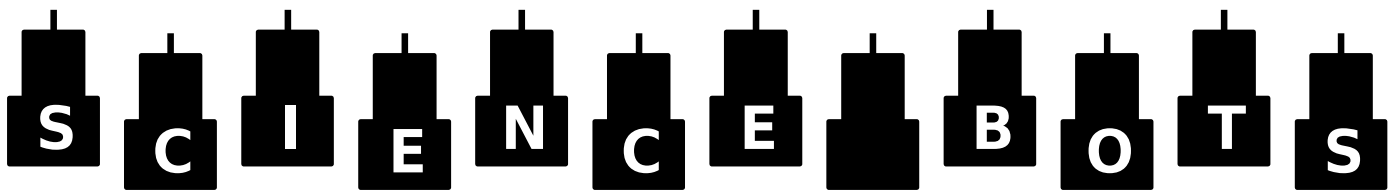
```
# Download data:
np get -c RAR5dwELYLKGSfrOclnWhjOj-2nGZN_8BW1JjxwFZINHw
# Analyze data:
...
```

Existing data can be recombined in new indexes; and researchers can unambiguously refer to the used datasets for new results:

```
this: prov:wasDerivedFrom nps:RAR5dwELYLKGSfrOclnWhjOj-2nGZN_8BW1Jjx
```

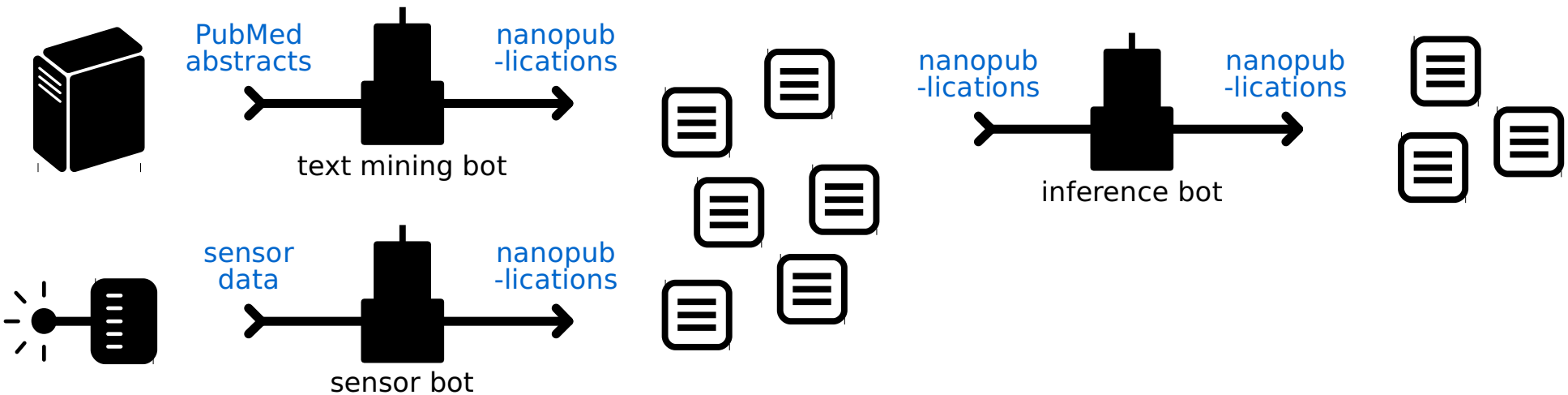
Kuhn et al. Publishing without Publishers: a Decentralized Approach to Dissemination, Retrieval, and Archiving of Data. arXiv:1411.2749.

Could these techniques and infrastructures allow us to make a step forward in terms of automation in science?



Science Bots — Scientists' Little Helpers in the Future?

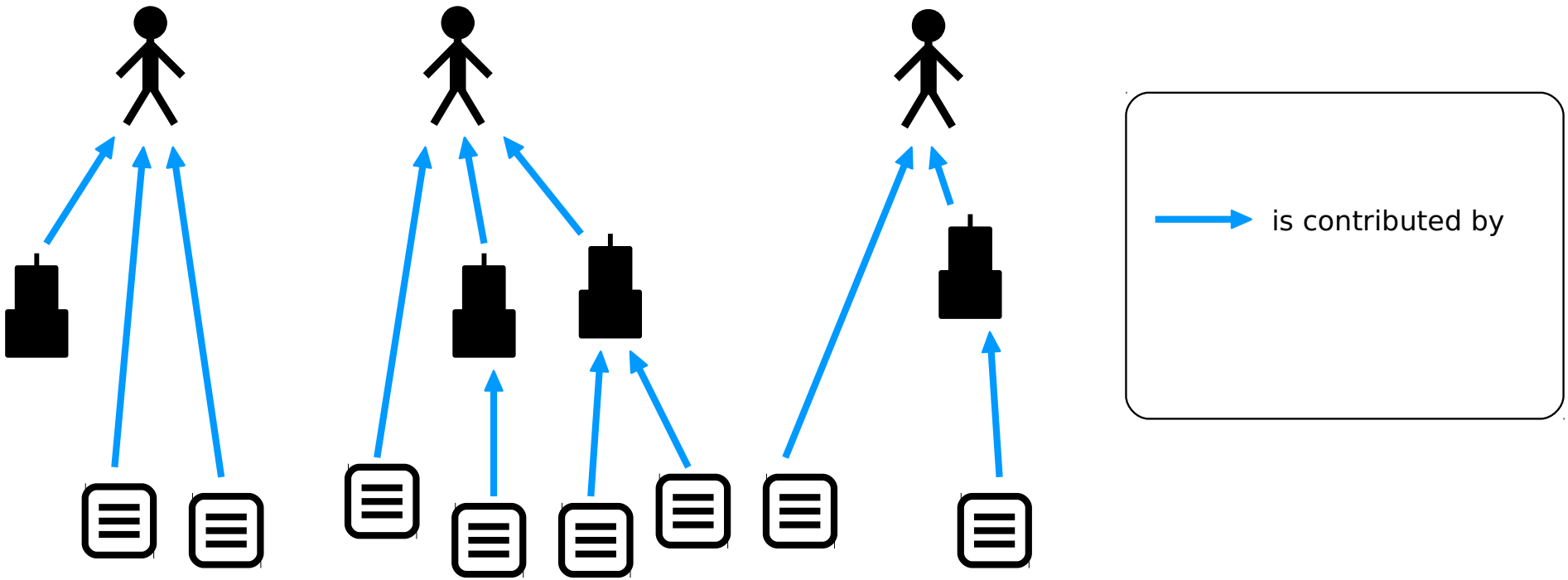
"Science bots" that autonomously publish results in their own name could cover a wide variety of applications, for example:



Kuhn. Science Bots: A Model for the Future of Scientific Computation? SAVE-SD, WWW 2015 Companion Proceedings.

Quality Control with Reputation Mechanisms and Network Metrics?

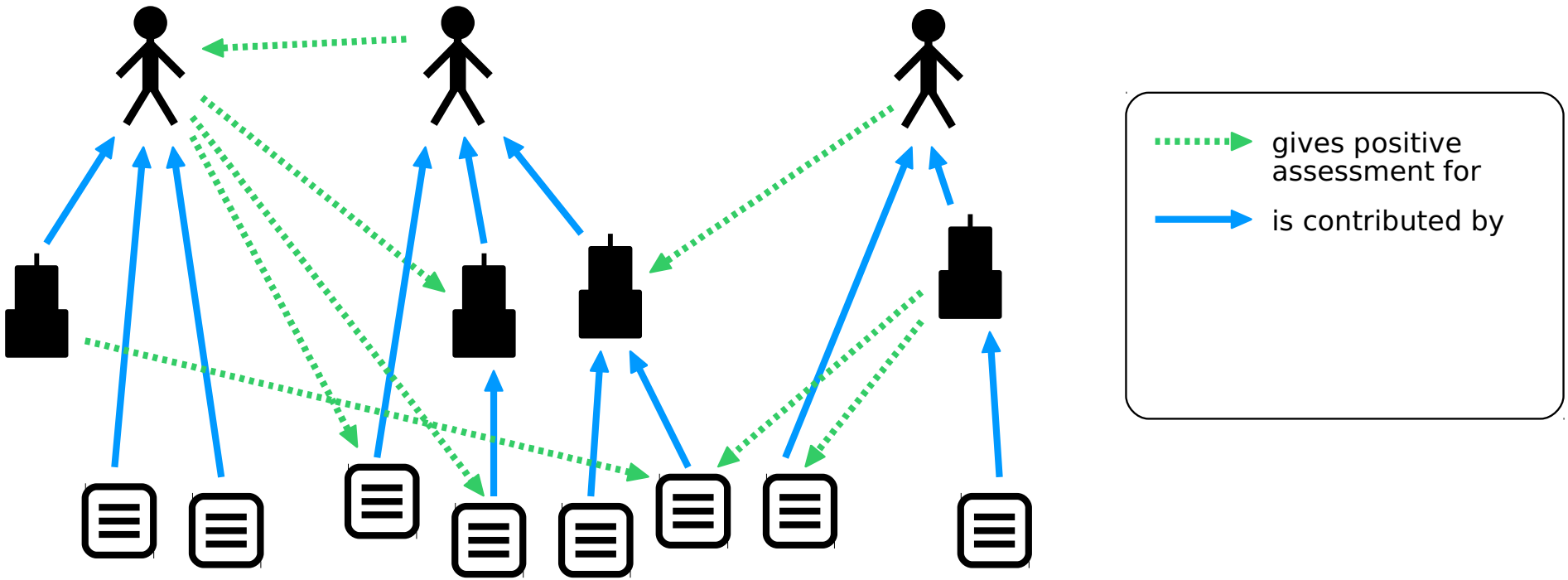
Robust automatic calculation of reputation metrics in a decentralized and open system:



Kuhn. Science Bots: A Model for the Future of Scientific Computation? SAVE-SD, WWW 2015 Companion Proceedings.

Quality Control with Reputation Mechanisms and Network Metrics?

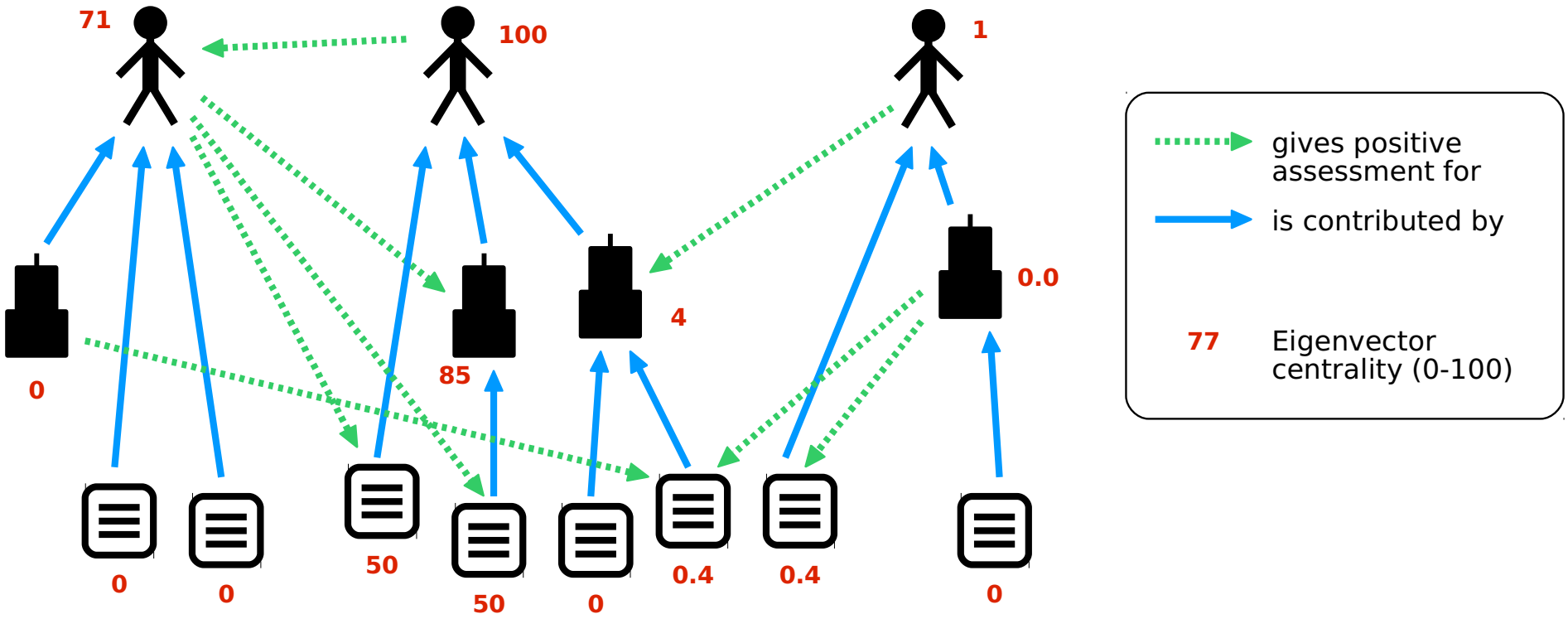
Robust automatic calculation of reputation metrics in a decentralized and open system:



Kuhn. Science Bots: A Model for the Future of Scientific Computation? SAVE-SD, WWW 2015 Companion Proceedings.

Quality Control with Reputation Mechanisms and Network Metrics?

Robust automatic calculation of reputation metrics in a decentralized and open system:



Kuhn. Science Bots: A Model for the Future of Scientific Computation? SAVE-SD, WWW 2015 Companion Proceedings.

Thank you for your attention!

Questions?

Further information:

- Trusty URIs: http://trustyuri.net
- Nanopublications: http://nanopub.org
- Nanopublication Server Network: http://arxiv.org/abs/1411.2749
- Science Bots: http://arxiv.org/abs/1503.04374