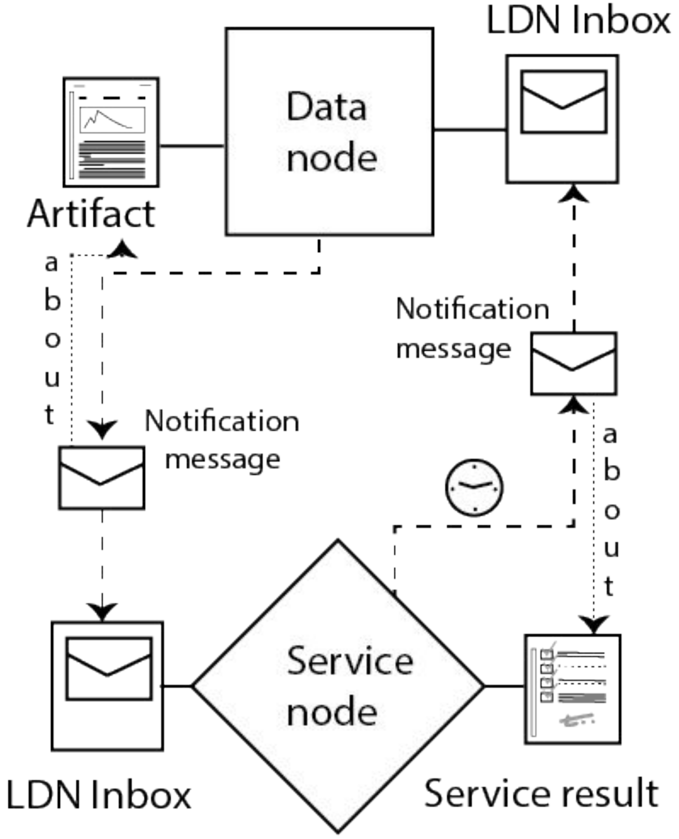
Decentralized Scholarly Communication & the Notify Protocol

Patrick Hochstenbach (UGent) &
Herbert Van de Sompel (DANS & UGent)



Researcher Pod: Project Overview

- PIs: Ruben Verborgh, University of Ghent
- Original PIs: Ruben Verborgh & Herbert Van de Sompel (DANS)
- PhD students: Ruben Dedecker, **Patrick Hochstenbach**
- Duration: 01/01/2020-31/12/2023
- Funding:
  - Andrew W. Mellon Foundation
  - ± $ 800K (staff & travel)
- Topic: Technical aspects of a decentralized, decoupled scholarly communication system
  - Inspired by my 2017 CNI Paul Evan Peters lecture “Scholarly Communication: Deconstruct & Decentralize?”, see https://www.youtube.com/watch?v=o4nUe-6Ln-8

Researcher Pod Project: Combining Two Perspectives

1. **Decoupled Scholarly Communication System**

2. Decentralized Web

Functions of Scholarly Communication

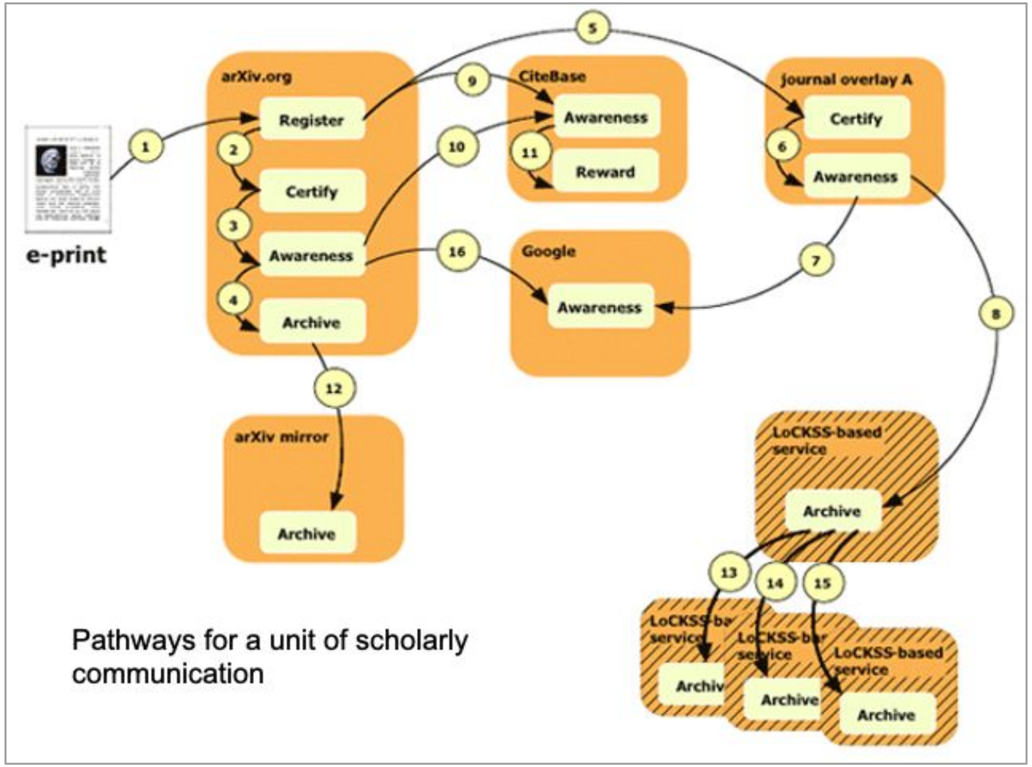
- Registration: Allows claims of precedence for a scholarly finding
- Certification: Establishes validity of the claim
- Awareness: Allows actors in the system to remain aware of new claims
- Archiving: Preserves the scholarly record over time

Roosendaal and Geurts (1997) Forces and Functions in Scientific Communication.
http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html

Decoupling the Functions

- In a digital networked scholarly communication system:
  - Each function can be fulfilled by a different party
  - Each function can be fulfilled in different ways
  - Each function can simultaneously be fulfilled by different parties, potentially in different ways

Research Outputs go through a Value Chain

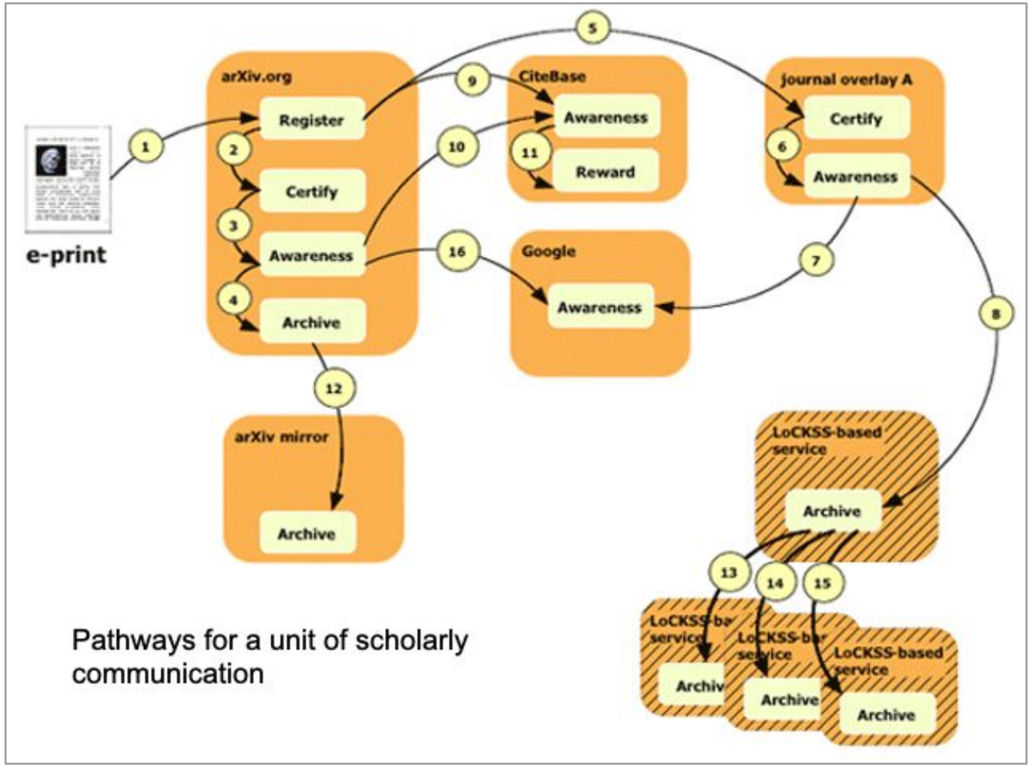


Pathways for a unit of scholarly communication

Van de Sompel, H., et al. (2004) Rethinking scholarly communication: Building the System that Scholars Deserve. D-Lib Magazine, 10(9). https://doi.org/10.1045/september2004-vandesompel

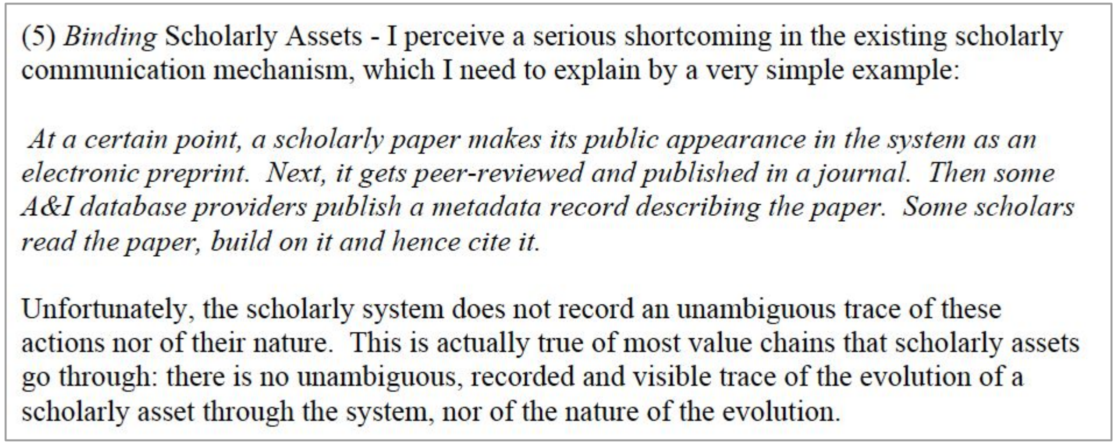
Research Outputs go through a Value Chain

In order for this to be realistically feasible/scalable, interoperability needs to be established for communication with the parties/services that fulfill the functions



Van de Sompel, H., et al. (2004) Rethinking scholarly communication: Building the System that Scholars Deserve. D-Lib Magazine, 10(9). https://doi.org/10.1045/september2004-vandesompel

Record/Expose Value Adding Events


(5) *Binding* Scholarly Assets - I perceive a serious shortcoming in the existing scholarly communication mechanism, which I need to explain by a very simple example:

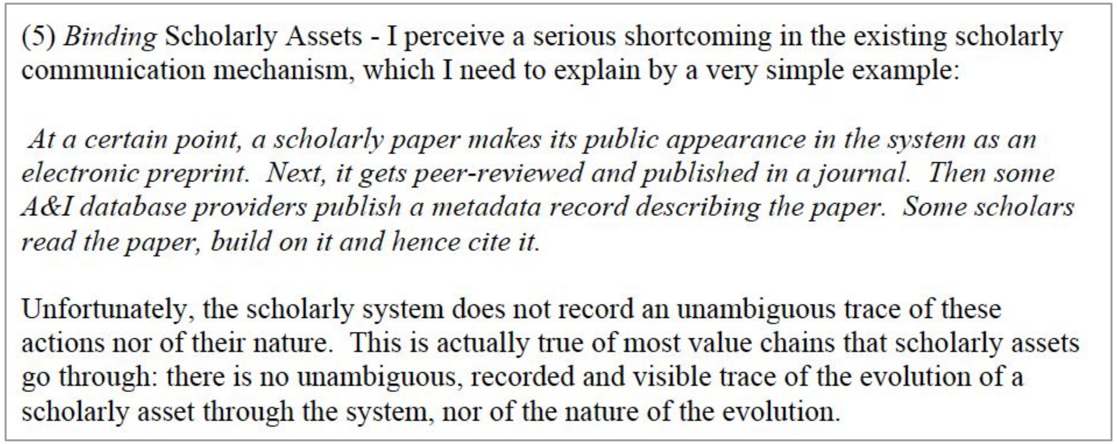
At a certain point, a scholarly paper makes its public appearance in the system as an electronic preprint. Next, it gets peer-reviewed and published in a journal. Then some A&I database providers publish a metadata record describing the paper. Some scholars read the paper, build on it and hence cite it.

Unfortunately, the scholarly system does not record an unambiguous trace of these actions nor of their nature. This is actually true of most value chains that scholarly assets go through: there is no unambiguous, recorded and visible trace of the evolution of a scholarly asset through the system, nor of the nature of the evolution.

Van de Sompel, H. (2003) Roadblocks. NSF Workshop Post Digital Library Futures
https://web.archive.org/web/20031020185620/http://www.sis.pitt.edu/~dlwkshop/paper_sompel.html

Record/Expose Value Adding Events

In order to achieve this, value added events need to be recorded, uniformly published, and made discoverable.



(5) *Binding* Scholarly Assets - I perceive a serious shortcoming in the existing scholarly communication mechanism, which I need to explain by a very simple example:

At a certain point, a scholarly paper makes its public appearance in the system as an electronic preprint. Next, it gets peer-reviewed and published in a journal. Then some A&I database providers publish a metadata record describing the paper. Some scholars read the paper, build on it and hence cite it.

Unfortunately, the scholarly system does not record an unambiguous trace of these actions nor of their nature. This is actually true of most value chains that scholarly assets go through: there is no unambiguous, recorded and visible trace of the evolution of a scholarly asset through the system, nor of the nature of the evolution.

Van de Sompel, H. (2003) Roadblocks. NSF Workshop Post Digital Library Futures
https://web.archive.org/web/20031020185620/http://www.sis.pitt.edu/~dlwkshop/paper_sompel.html

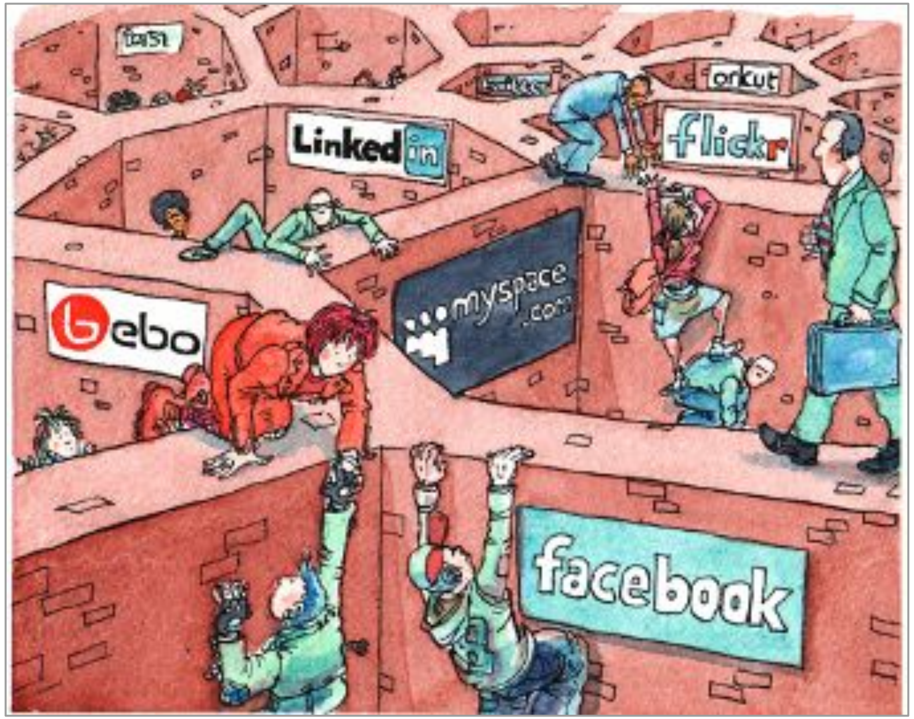
Researcher Pod Project: Combining Two Perspectives

1. Decoupled Scholarly Communication System

2. **Decentralized Web**

Centralized Web

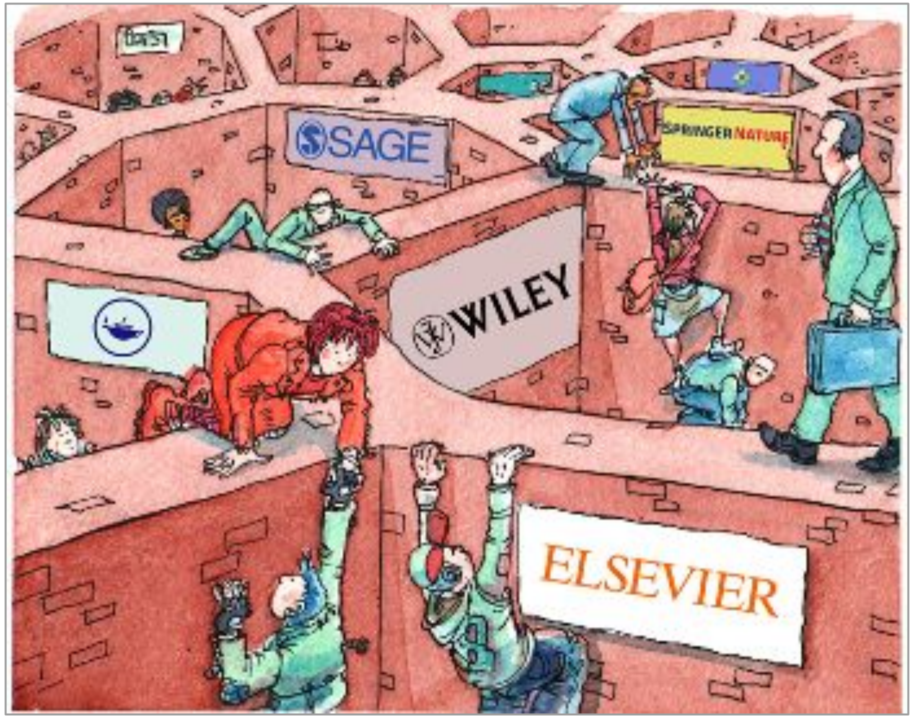
- Some massive central portals dominate the web
- Service is smooth, free; the user is the product
- No interoperability; different APIs for different platforms
- Functionality contained within a portal, can't be reused on content that resides in other portal



Tim Berners-Lee (2011) Socially aware cloud storage.
https://www.w3.org/DesignIssues/CloudStorage.html

Centralized Scholarly Communication

- Some massive publishers dominate scholcomm
- Consolidation of tools that span the research lifecycle
- Surveillance, data analytics
- Interoperability typically provided through central approaches
  - Central parties become too important to fail



Van de Sompel H. (2017) Scholarly Communication: Deconstruct & Decentralize?. https://www.youtube.com/watch?v=o4nUe-6Ln-8

Researcher Pod Project

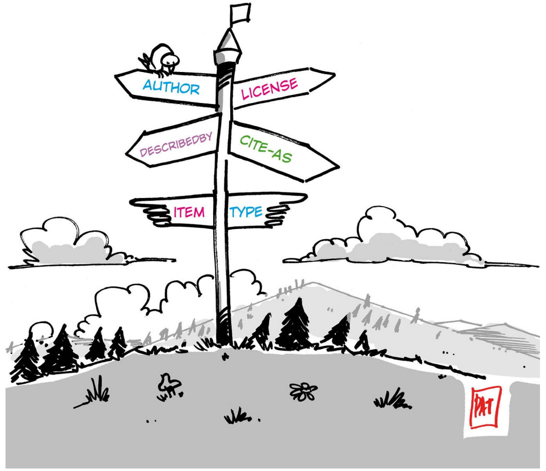
As such, the project is exploring solutions to existing problems that:

- support decoupling the functions of scholarly communication
    - level the playing field for new entrants
    - invite creativity

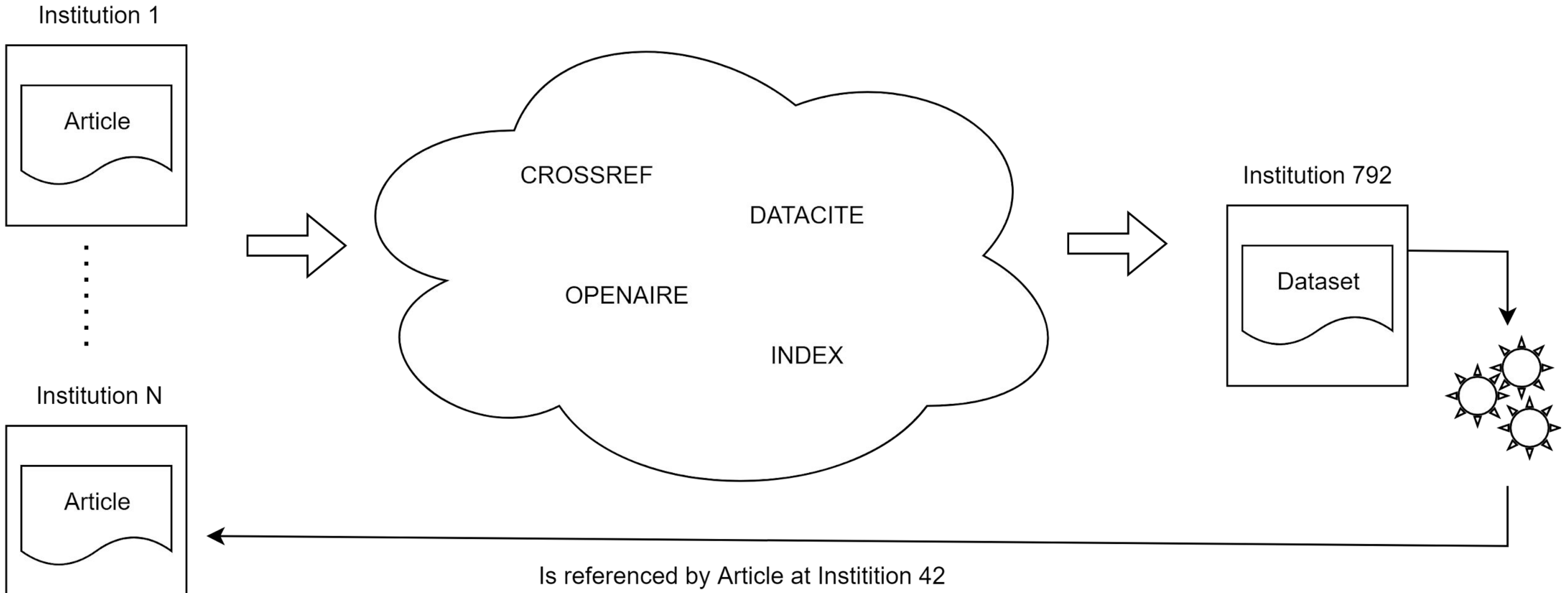
- do not require central parties
    - avoid “too big to trust” and “too important to fail”
    - repositories as starting points of value chains
        - cf. COAR Next Generation Repositories

Patrick Hochstenbach

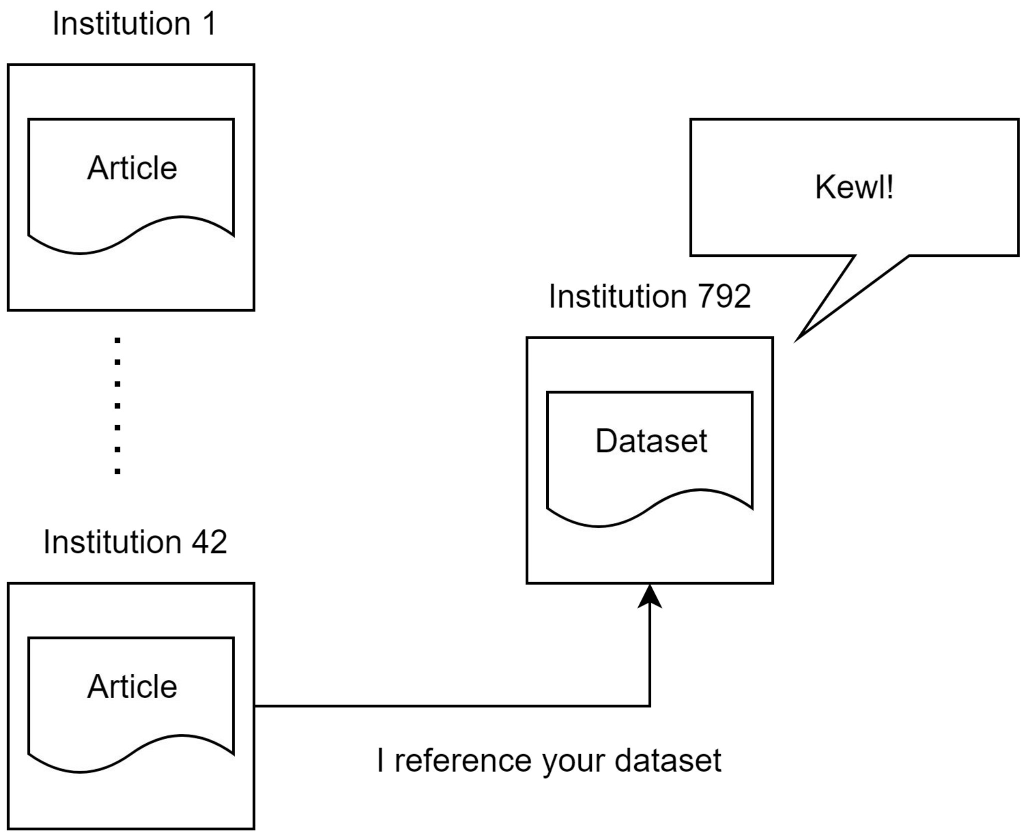
- Was on my team at:
  - UGent Library IT
  - LANL Prototyping Team
- Played a crucial role in my PhD research
  - development of SFX linking server prototypes
  - support of large-scale experiments
  - co-author on all papers
- Highly regarded contributor to library IT
- Talented illustrator



Problem statement - actual



Problem statement - counterfactual

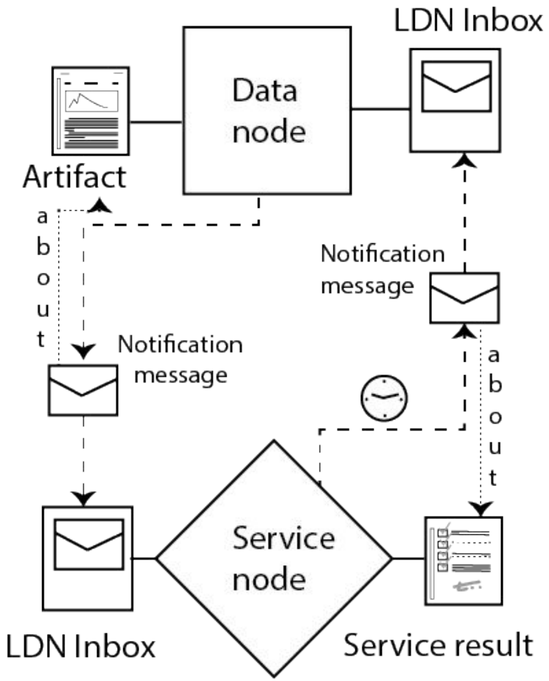


Need of decentralization & decoupling

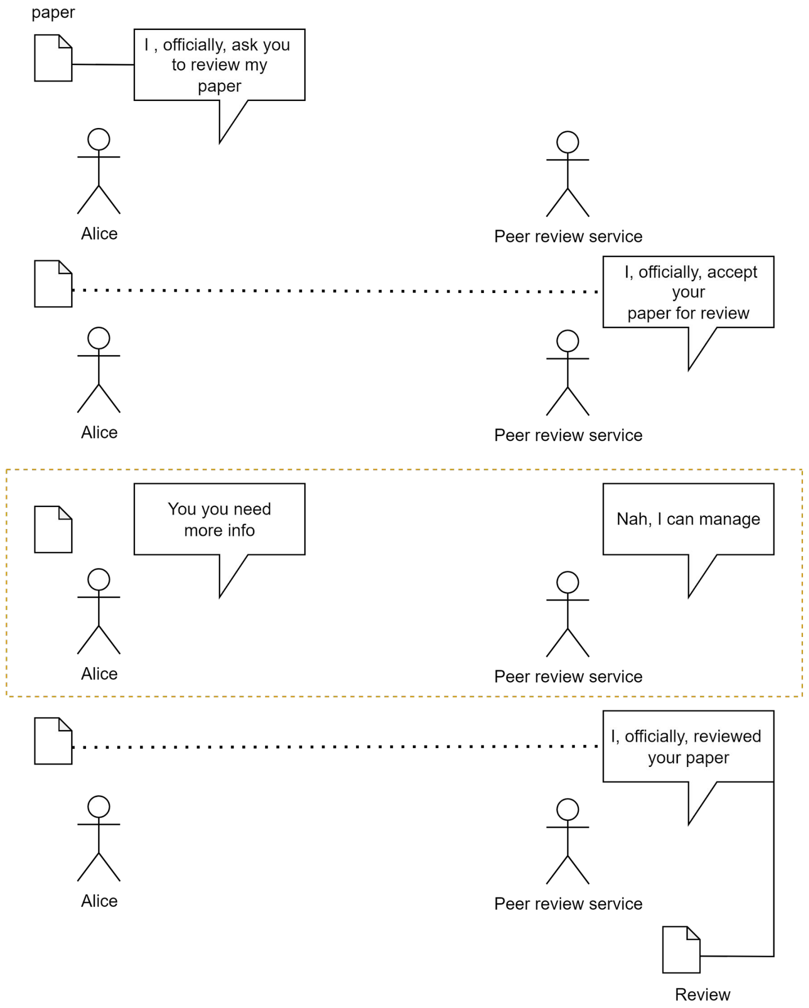
- The **value-chain** of what happens to (scholarly) artifacts is not written down explicitly
  - When were artifacts *registered*, *peer reviewed*, *published*, *indexed*, *archived*, *linked*?
- **This information is available** in the network at the moment it happens but it is **not disseminated** and often even **not stored**
- Currently need **post-factum** harvesting, indexing, processing to gather all this information
- This processing can only be done by **a happy few**
- Those that do this for free, can **stop providing the service** whenever they want (MS Academic Graph)
- Those that get funded/payed, become **too big to fail**

Introducing the Value-Added network

- Network of Data Nodes and Service Nodes with read-write capabilities
- That keep each-other in touch about important value-adding lifecycle events
  - Registration
  - Certification
  - Awareness
  - Archiving
  - + Linking



Value-adding?



Value Adding

Make machine readable!

Value Adding

Make machine readable!

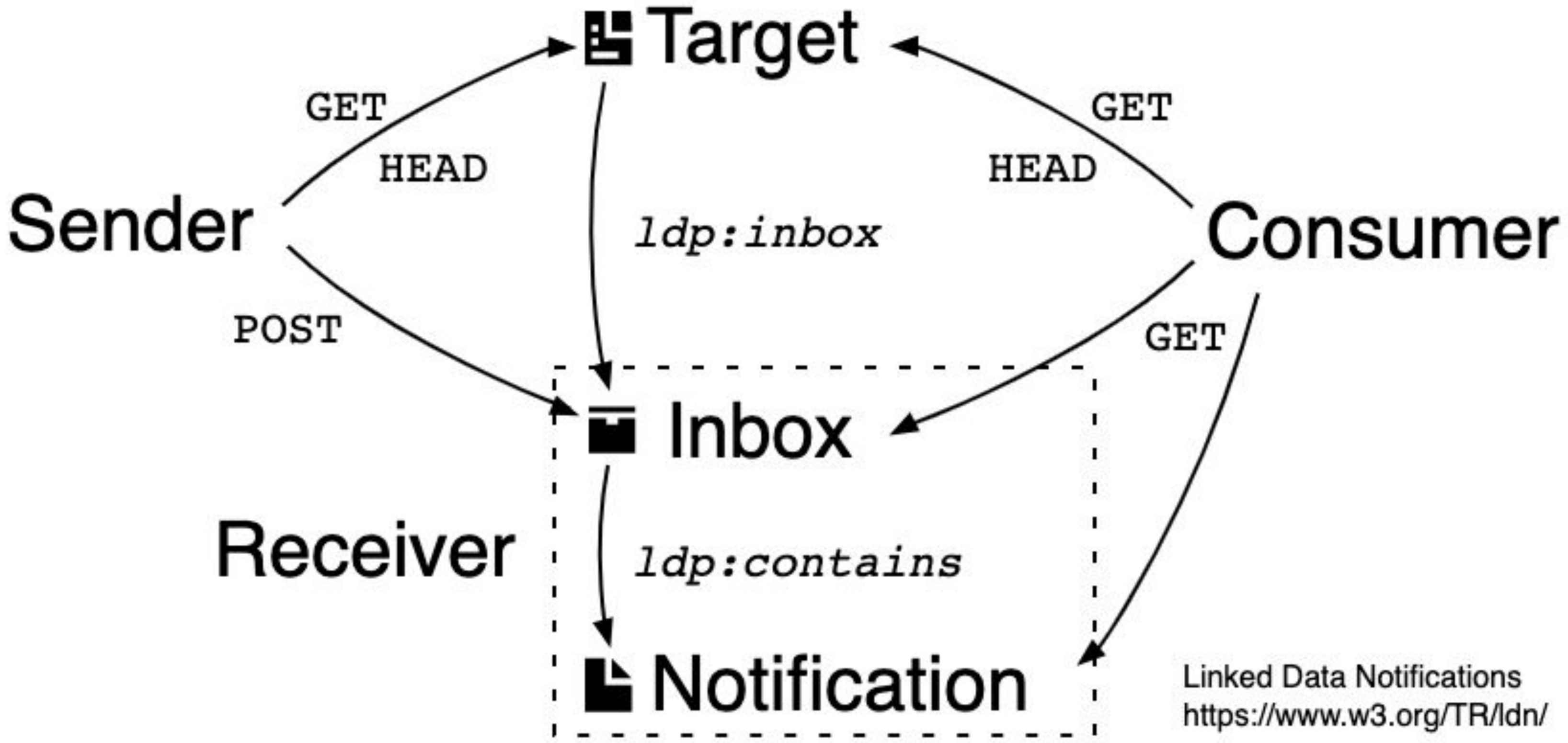
Value Adding

Make machine readable!

Technologies

- Linked Data Notifications (LDN)
- Activity Streams 2.0 (AS)
- Related to projects such as:
  - Dokieli (Carven Capadisli)
  - ActivityPub (Mastodon, Peertube, …)
  - COAR Notify
  - Researcher Pod
  - ErfgoedPOD
  - DICE DDPS

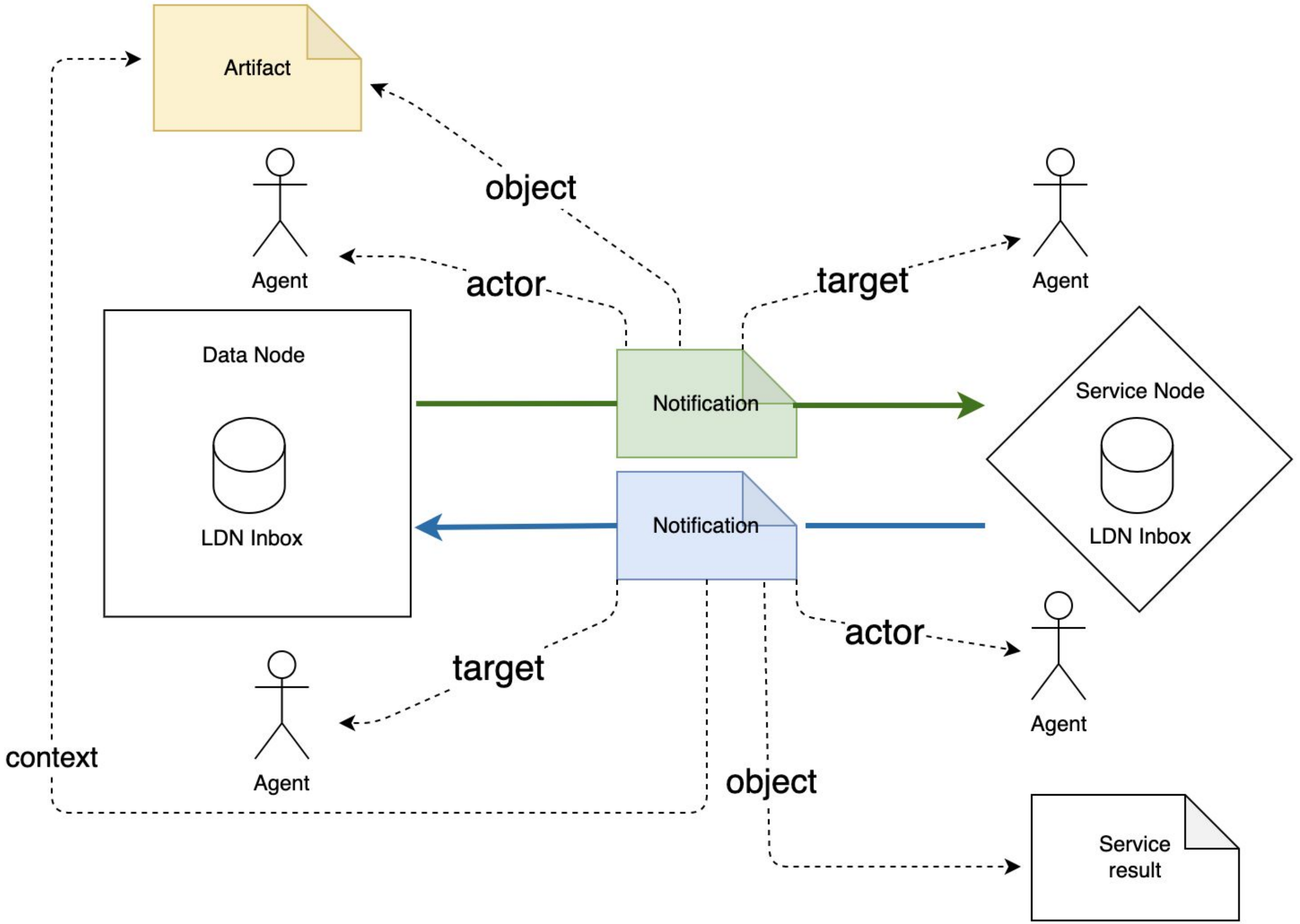
Linked Data Notifications



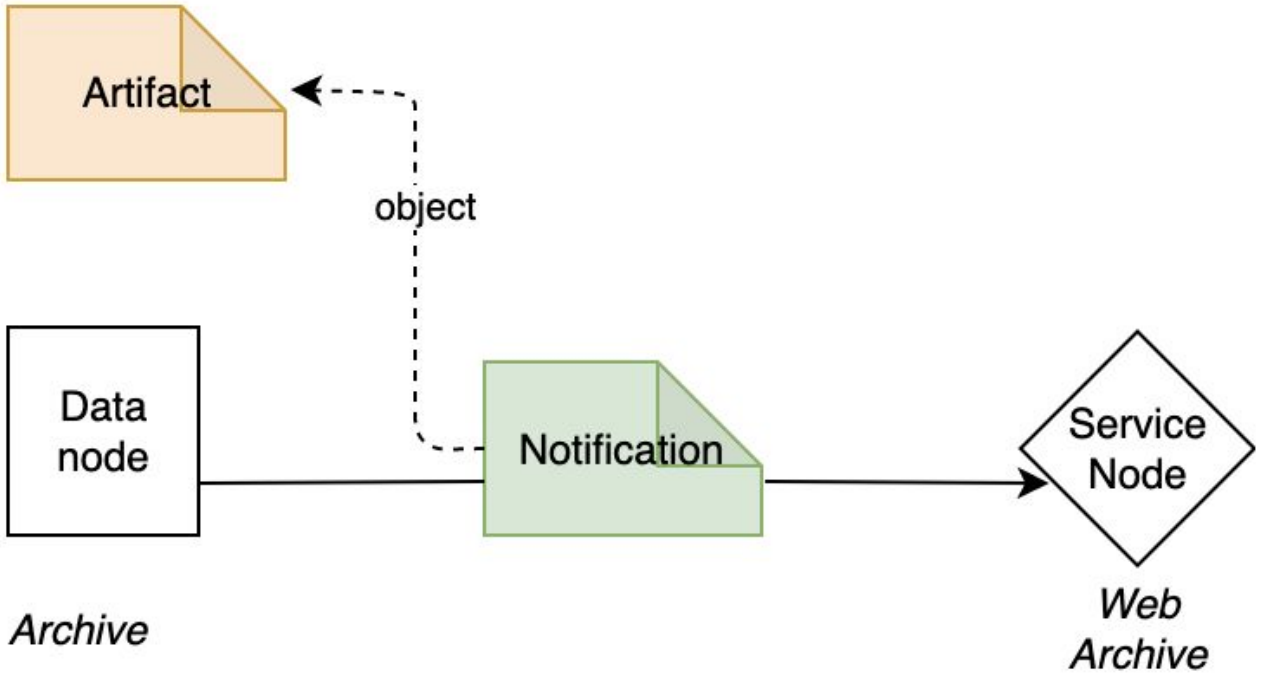
Overview of Linked Data Notifications

We introduce new terminology

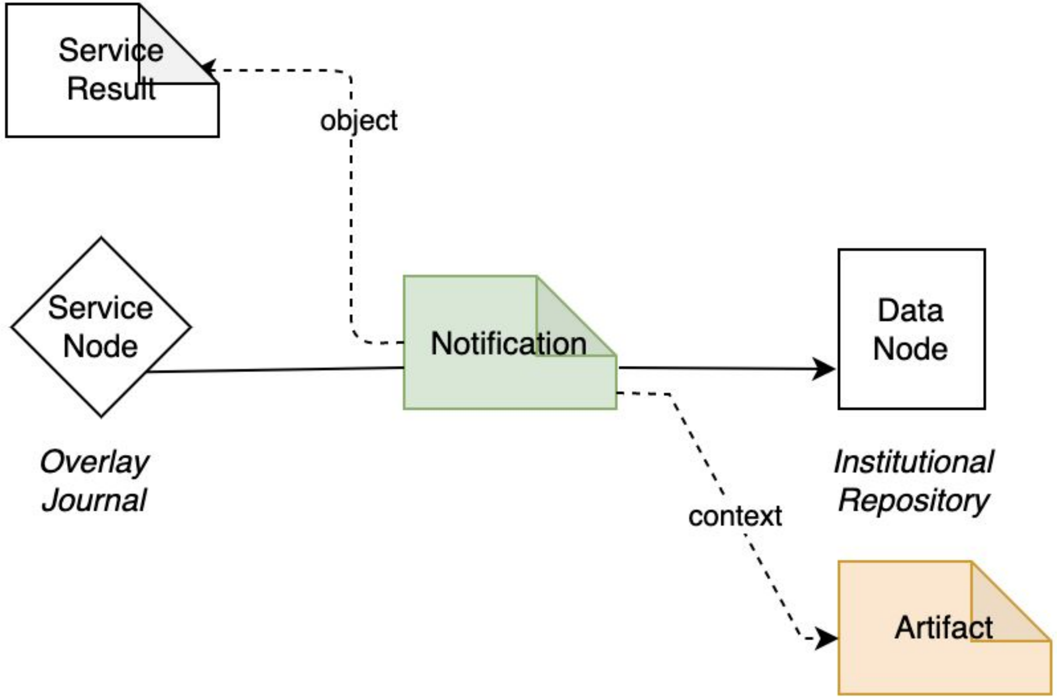
- *Artifact*
  - LDN Target = landing page + optional [hypermedia controls, linked data]
  - *Articles, Books, Datasets, Software, … , Part of scholarly record*
- *Service Result*
  - Result of a value adding service
  - *Peer Review, Memento, Indexed Webpage, Link Description, …*
- *Agent* (A)
  - LDN Sender | Consumer of LDN Notifications
- *Data Node* (DN)
  - A LDN Target that hosts *Artifacts*
  - Provides as LDN Receiver one or more LDN *Inboxes* for these artifacts
- *Service Node* (SN)
  - A LDN Target that produces *Service Results*
  - Provides as LDN Receiver one or more LDN Inboxes that trigger services


Artifact
object
Agent
actor
target
Agent
Data Node
LDN Inbox
Notification
Service Node
LDN Inbox
Notification
actor
Agent
target
Agent
context
object
Service result

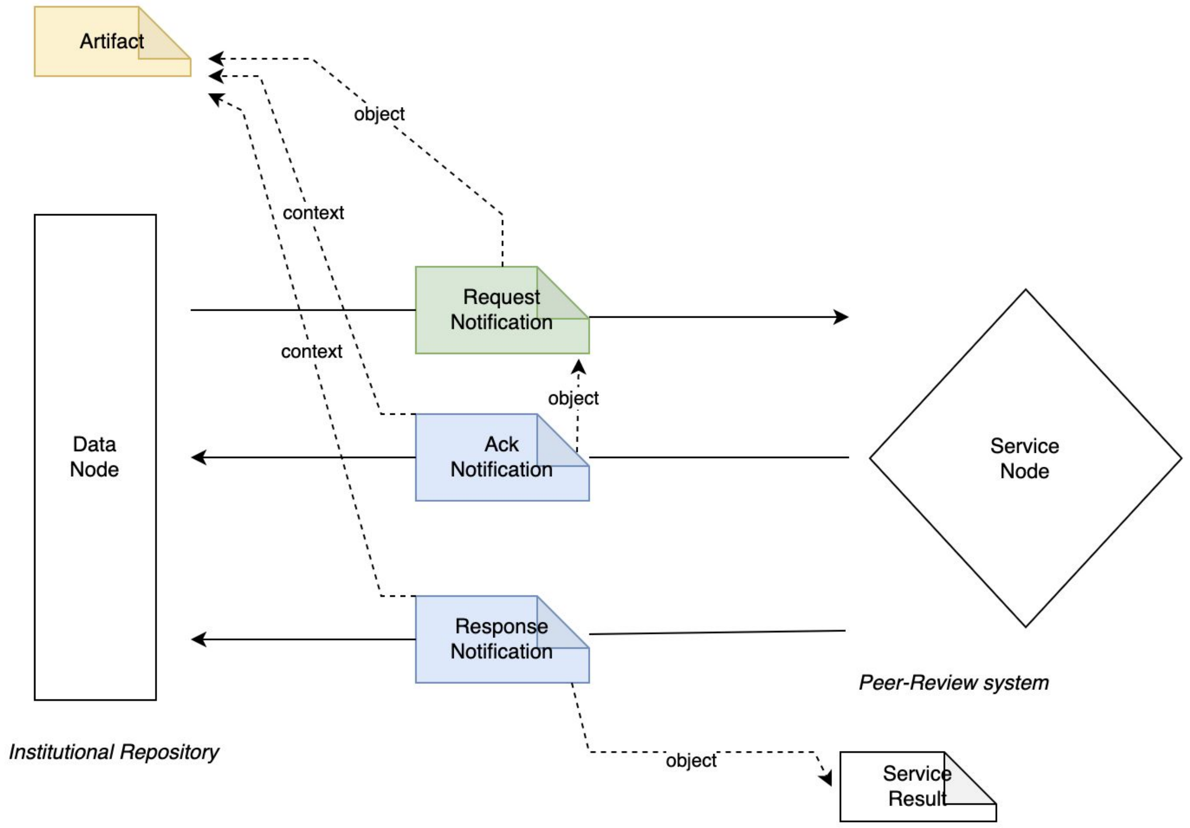
Network communication patterns : one-way



Network communication patterns : one-way



Network communication patterns : request-response



Activity Streams 2.0

Anatomy of a notification message (I)

- Notifications have an Activity Streams 2.0 (`as:`) payload
- Default serialization is JSON-LD (but other serializations are also allowed)
- Every payload must have one or more rdf:type properties
  - At least one is a subtype of `as:Activity` that is part of the subset:
    1. as:Announce, as:Offer, as:Accept, as:Reject, as:Undo
    2. as:Create, as:Update, as:Remove
- Subset i is for activities about value adding life-cycle events
- Subset ii is for activities about CRUD life-cycle events
- Communities can introduce new subtypes:
  - E.g. `coar-notify:ReviewAction`

Anatomy of a notification message (II)

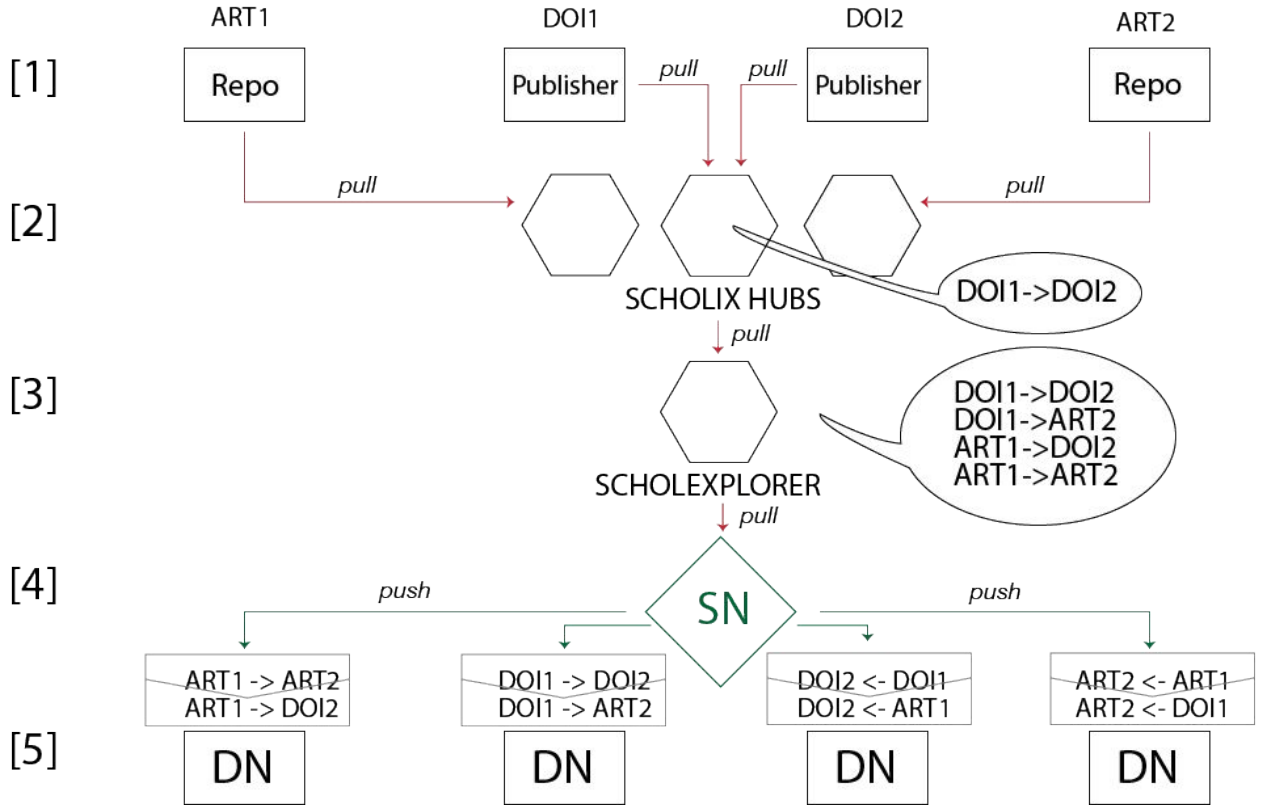
AS2 element	Description
`id`	Message identifier
`type`	Activity type
`as:actor`	Agent that performed the activity
`as:origin`	Agent responsible for sending the notification
`as:context`	The artifact on the data node for which an value-added service was provided
`as:object`	The result of the value-added service provided for an artifact on the data node
`as:target`	The agent at the data node that is the addressee of the notification

Anatomy of a one-way pattern from a service node to a data node

```
@prefix as: <https://www.w3.org/ns/activitystreams#> .
@prefix ldp: <http://www.w3.org/ns/ldp#> .

<urn:uuid:239FD510-03F4-4B56-B3A0-0D3B92F3826D> a as:Announce ;
  as:actor <https://fairfield.org/about#us> ;
  as:origin <https://fairfield.org/system> ;
  as:context <https://springfield.library.net/artifact/13-02.html> ;
  as:object <urn:uuid:CF21A499-1BDD-4B59-984A-FC94CF6FBA86> ;
  as:target <https://springfield.library.net/about#us> .

<https://fairfield.org/about#us> a as:Organization ;
  ldp:inbox <https://fairfield.org/inbox> ;
  as:name "Fairfield Archive" .

<https://fairfield.org/system> a as:Service ;
  as:name "Fairfield Archive System" .

<urn:uuid:CF21A499-1BDD-4B59-984A-FC94CF6FBA86> a as:Relationship ;
  as:subject <https://springfield.library.net/artifact/13-02.html> ;
  as:relationship <https://www.iana.org/memento> ;
  as:object <https://fairfield.org/archive/version/317831-13210> .

<https://springfield.library.net/about#us> a as:Organization ;
  ldp:inbox <https://springfield.library.net/inbox/> ;
  as:name "Springfield Library" .
```

Experiment

Scholix Framework



Scholix message

Source		
	{ …metadata…}	values
	Identifiers	pmc:PMC7005061 pmid:32029780 handle:1854/LU-8646849 …
Relationship	"References"	
Target		
	{...metadata…}	values
	Identifiers	doi:10.5061/dryad.10hq7 …

Generating notification messages I

a		as:Announce
as:actor		https://scholexplorer.openaire.eu/#about
as:origin		https://mellonscholarlycommunication.github.io/about#us
as:context		{ an artifact on a data node }
as:object		
	a	as:Relationship
	as:subject	{ a subject URI }
	as:relationship	{ a relationship URI}
	as:object	{ an object URI}
as:target		{ the target node + LDN inbox }

Scholix message

Source		
	{ …metadata…}	values
		pmc:PMC7005061 pmid:32029780 handle:1854/LU-8646849 …
Relationship	"References"	
Target		
	{...metadata…}	values
	Identifiers	doi:10.5061/dryad.10hq7 …

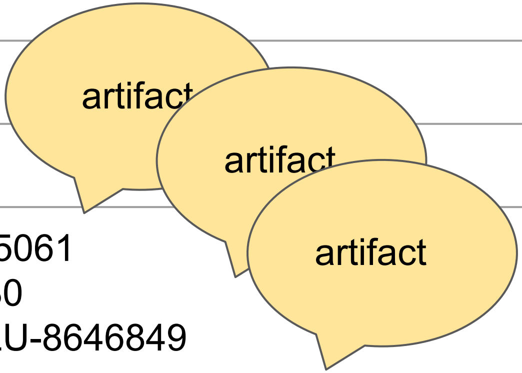
as:subject

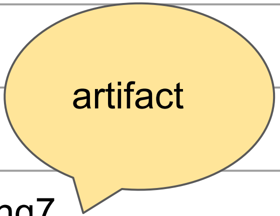
as:relationship

as:object

Scholix message

Source		
	{ …metadata…}	values
	Identifiers	pmc:PMC7005061 pmid:32029780 handle:1854/LU-8646849 …
Relationship	"References"	
Target		
	{...metadata…}	values
	Identifiers	doi:10.5061/dryad.10hq7 …





Discovery of LDN Inboxes

Algorithm:

- Deference a PID up to their landing pages (follow HTTP 302 redirects until a HTTP 200 can be found)
- Read the `http://www.w3.org/ns/ldp#inbox` relation from the HTTP Link headers
    - If found, then this is the Target LDN Inbox
    - Else (not part of our spec)
        - LDN Inbox := baseUrl(landing_page) + '/inbox'
        - E.g. `https://arxiv.org/abs/2204.03383` -> `https://arxiv.org/inbox`

Generating notification messages II

a		as:Announce
as:actor		https://scholexplorer.openaire.eu/#about
as:origin		https://mellonscholarlycommunication.github.io/about#us
as:context		https://biblio.ugent.be/publication/8159575
as:object		
	a	as:Relationship
	as:subject	https://biblio.ugent.be/publication/8159575
	as:relationship	http://www.scholix.org/References
	as:object	https://doi.org/10.3410/f.1098070.554047
as:target		https://biblio.ugent.be/inbox/

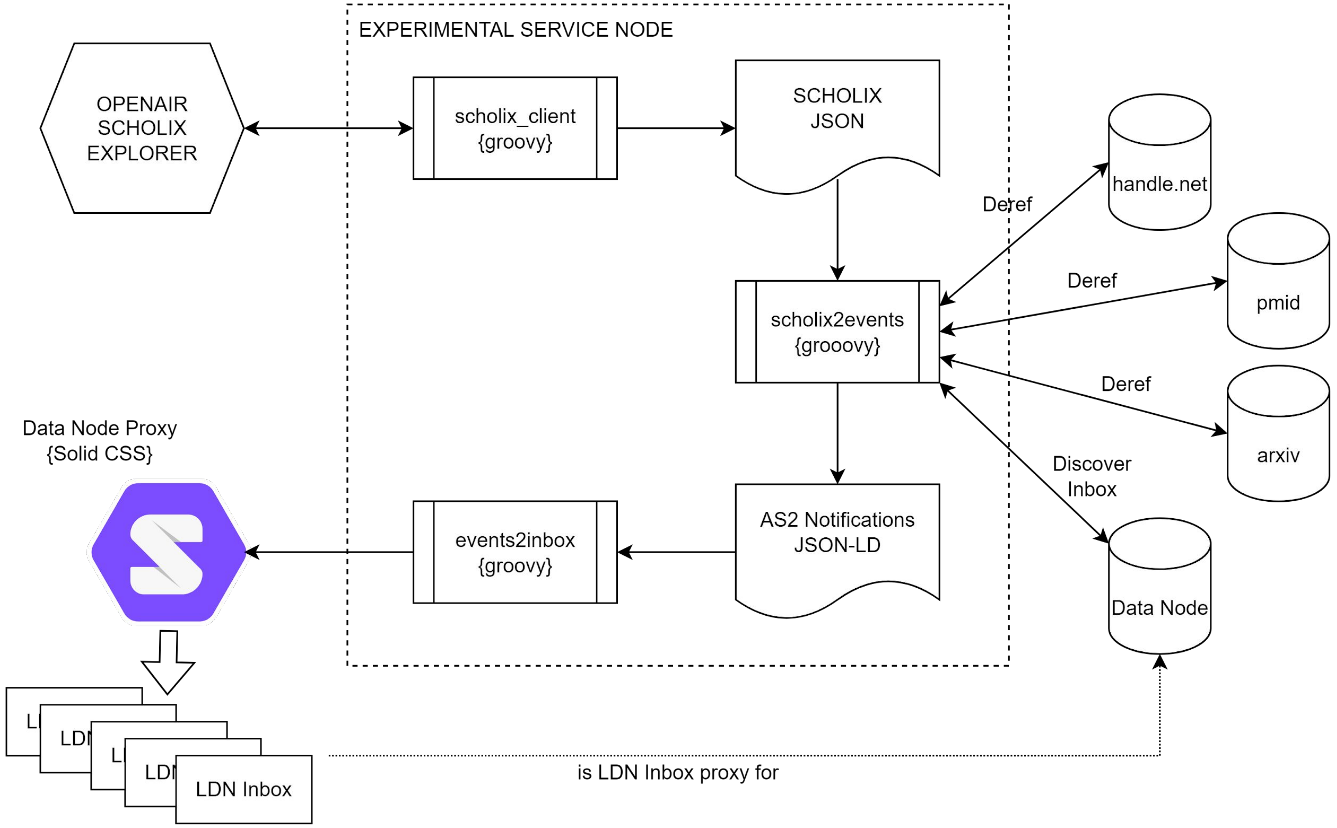
Sending notifications to LDN Inboxes

Source		
	{ …metadata…}	values
	Identifiers	pmc:PMC7005061 pmid:32029780 handle:1854/LU-8646849 …
Relationship	"References"	
Target		
	{...metadata…}	values
	Identifiers	doi:10.5061/dryad.10hq7 …

Send notifications to:

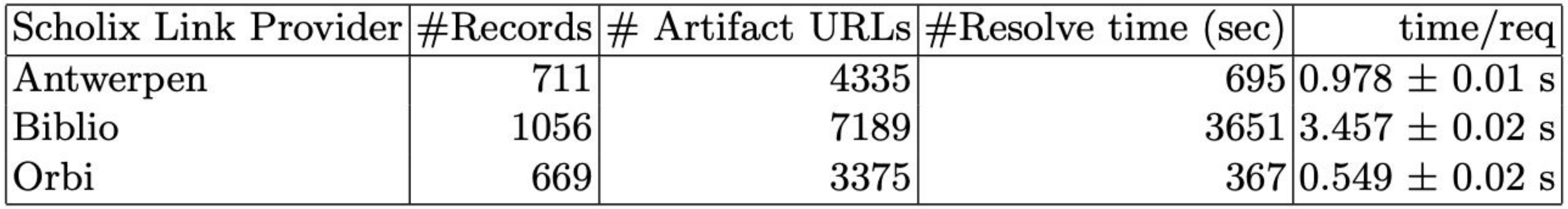
- pmc:PMC700561
  - pmc:... <-> doi:...
- pmid:32029780
  - pmid:... <-> doi:...
- handle:1864/LU-864689
  - handle:.. <-> doi:...
- doi:10.50601/dryiad.10hq7
  - pmc:... <-> doi:...
- doi:10.50601/dryiad.10hq7
  - pmid:... <-> doi:...
- doi:10.50601/dryiad.10hq7
  - handle:.. <-> doi:...

Implementation



Numbers

Table 2. Number of artifact URLs resolved for the data-literature network of each Belgian institution and time required to resolve PID-URLs to their landing page.



Scholix Link Provider	#Records	# Artifact URLs	#Resolve time (sec)	time/req
Antwerpen	711	4335	695	0.978 ± 0.01 s
Biblio	1056	7189	3651	3.457 ± 0.02 s
Orbi	669	3375	367	0.549 ± 0.02 s

Numbers

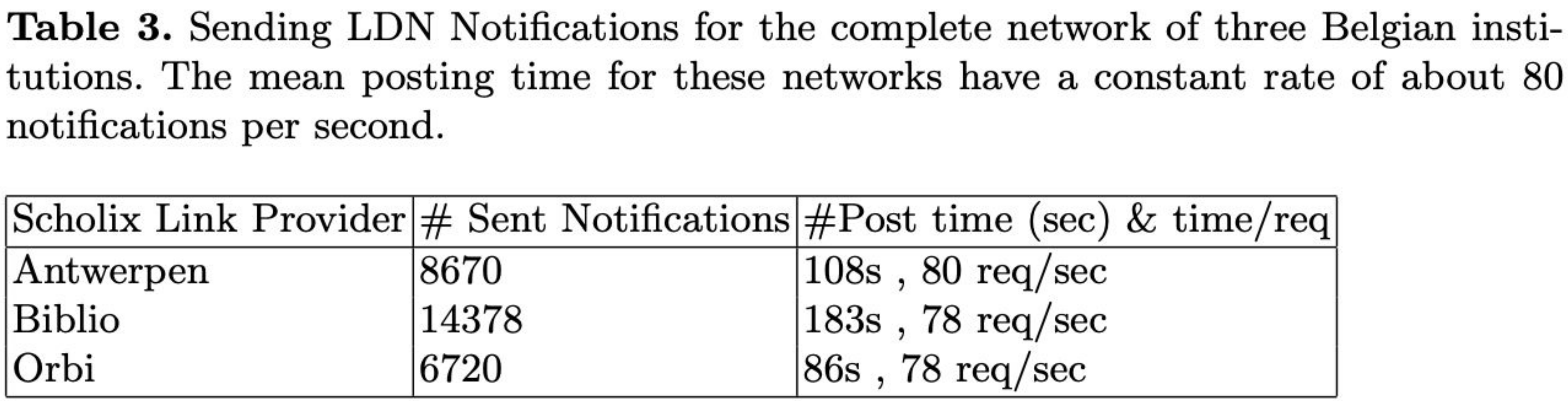


Table 3. Sending LDN Notifications for the complete network of three Belgian institutions. The mean posting time for these networks have a constant rate of about 80 notifications per second.

Scholix Link Provider	# Sent Notifications	#Post time (sec) & time/req
Antwerpen	8670	108s , 80 req/sec
Biblio	14378	183s , 78 req/sec
Orbi	6720	86s , 78 req/sec

Conclusion

- It is possible to add read-write capabilities on top of current research networks
- A demonstration was given how a national service node could distribute linking information to a network of Belgian repositories
- The scalability is dependent on the time it takes to resolve PID-urls, but even with our naive approach, the complexe Belgian Scholix linking information could be distributed within 2 hours on a small Linux host
- We are still dependent on data mining by Scholix and are creating experiments for direct communication between nodes
- Using Solid made implementing LDN Inboxes trivial