From Open Access to Open Standards, (Linked) Data and Collaborations

Simeon Warner https://orcid.org/0000-0002-7970-7855
(Director of IT for Library Linked Data and Repository Architecture, Cornell University Library, USA)

National Library of Finland Kirjastoverkkopäivät (Library Network Days), Helsinki, Finland, 2017-10-25

Tuesday 10 June
8:30am –6:00pm
Registration
Paasitorni, Entrance Hall
9:00am –9:30am
Welcome and introduction
Congress Hall
9:00am –6:00pm
Developer Challenge
Room 302-303
9:30am –10:45am
Keynote and discussion
Congress Hall

HPC
infrastructures
researchers
innovation

GIANT

simeon — mutt — 80×25

mutt

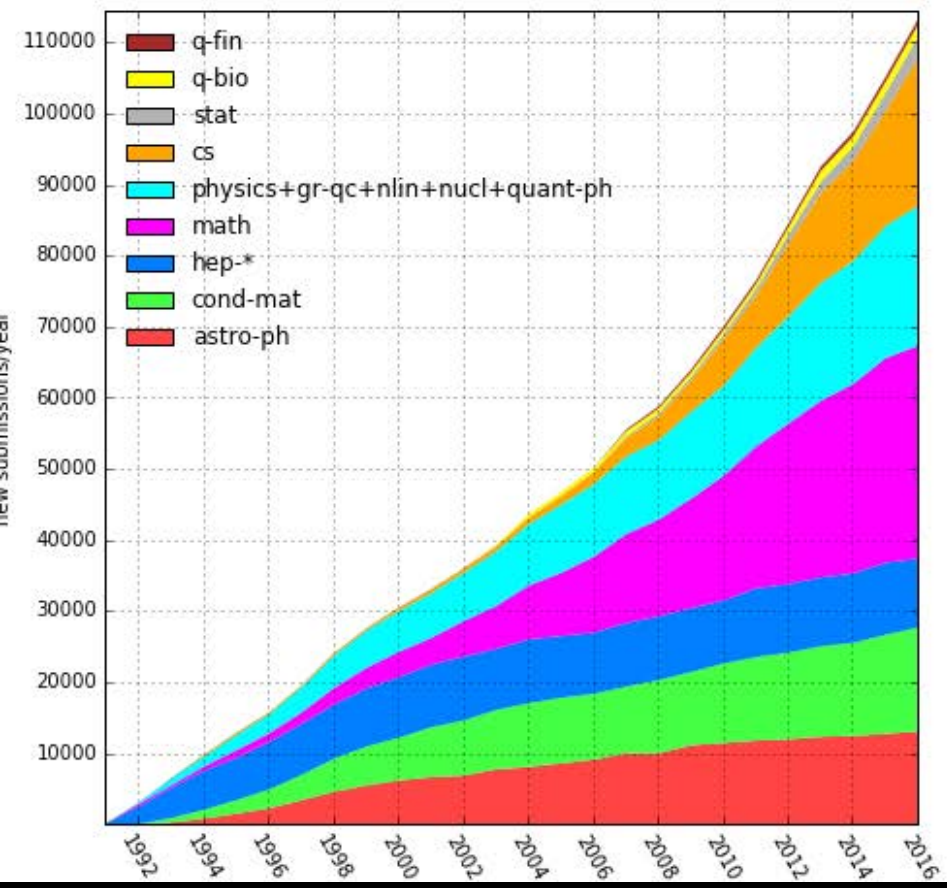
```
i:Exit  -:PrevPg  <Space>:NextPg v:View Attachm.  d:Del  r:Reply  j:Next ?:Help
Date: Thu, Aug 8 20:32:45 1996 -0500
From: send mail ONLY to hep-lat <no-reply@xxx.lanl.gov>
To: hep-lat daily title/abstract distribution <rabble@xxx.lanl.gov>
Subject: hep-lat daily Subj-class mailing f 151

------------------------------------------------------------------------------
 Submissions to: High Energy Physics - Lattice
 received from  Wed 07 Aug 96 00:00:02 GMT  to  Thu 08 Aug 96 00:00:04 GMT
------------------------------------------------------------------------------

\\
arXiv:hep-lat/9608042
From: Simon Catterall <sc@physics.syr.edu>
Date: Thu, 8 Aug 1996 16:12:33 GMT   (34kb)

Title: Simplicial Gravity in Dimension Greater than Two
Authors: S. Catterall, G. Thorleifsson, R. Renken and J. Kogut
Categories: hep-lat
Comments: Talk presented at LATTICE96(gravity)
Journal-ref: Nucl.Phys.Proc.Suppl. 53 (1997) 756-759
\\
  We consider two issues in the DT model of quantum gravity. First, it is shown
that the triangulation space for D>3 is dominated by triangulations containing
-N  - 1/25: send mail ONLY to he   hep-lat daily Subj-class mailing f 15 -- (6%)
```

How?

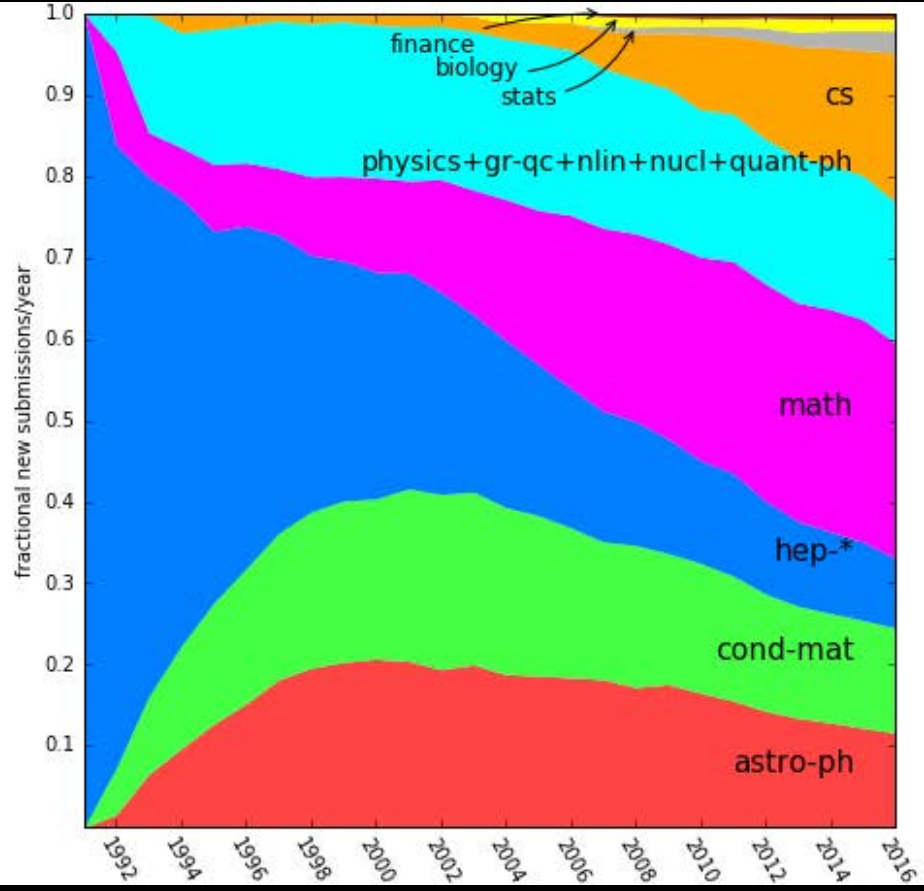
- This was xxx.lanl.gov, now known as arXiv.org
- I worked in a narrow field
- Everyone posted to one place
- It was a newish field
- (I was perhaps happy to not read widely enough)

https://arxiv.org/help/stats/2016_by_area/index

arXiv submissions



New submission rate,
color = subject



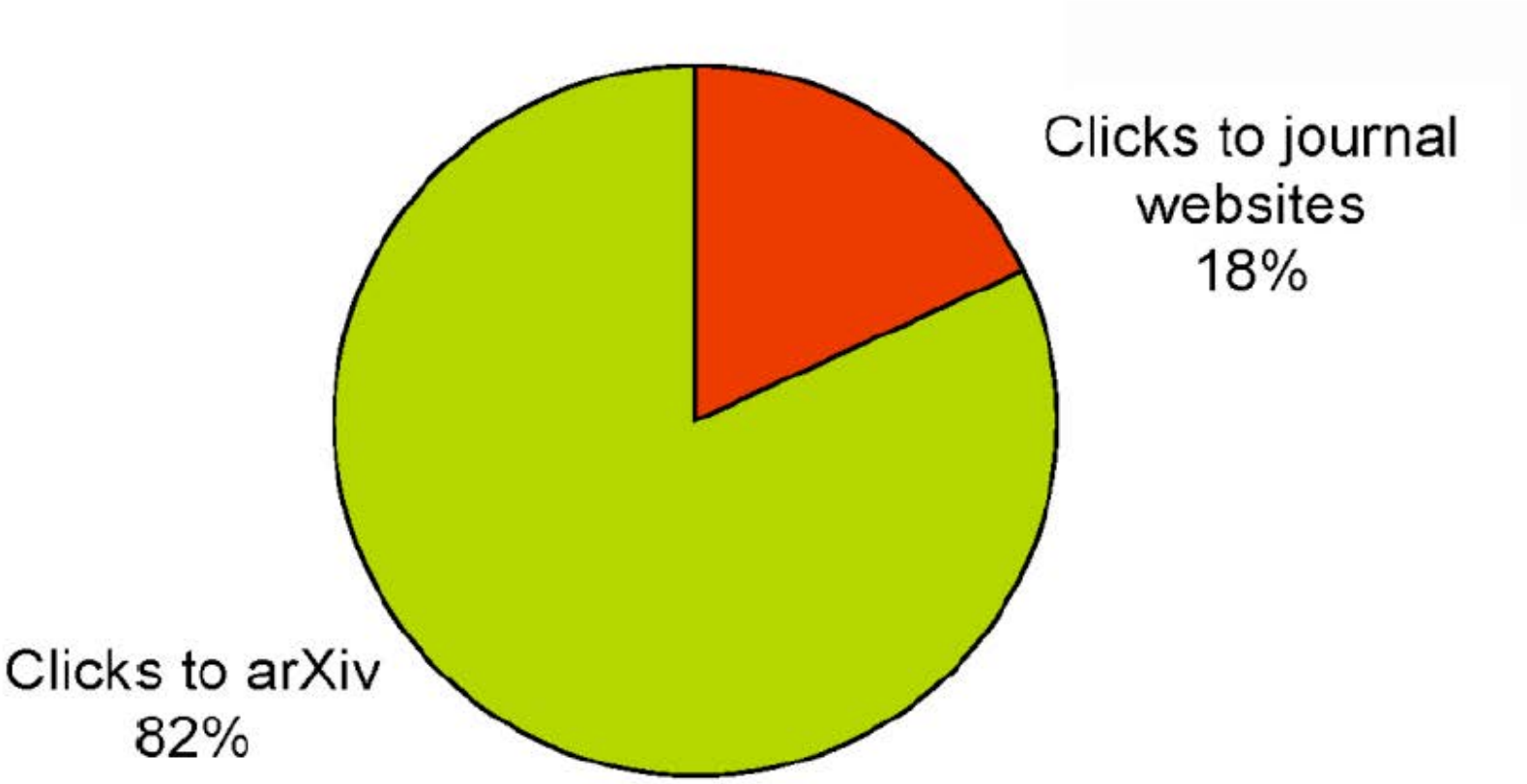
Fraction of total rate
for each subject area

What have we learned?

- Researchers are happy to use e-prints
- E-print repositories can scale
- Cost is low ($10-15/article)
- Some moderation necessary
- Not very disruptive to journal publishing (in physics)

Demonstrates substrate for article distribution supporting overlay, but there has not been significant adoption of overlap model

arxiv:0906.5418
Gentil-Beccot, Mele, Brooks



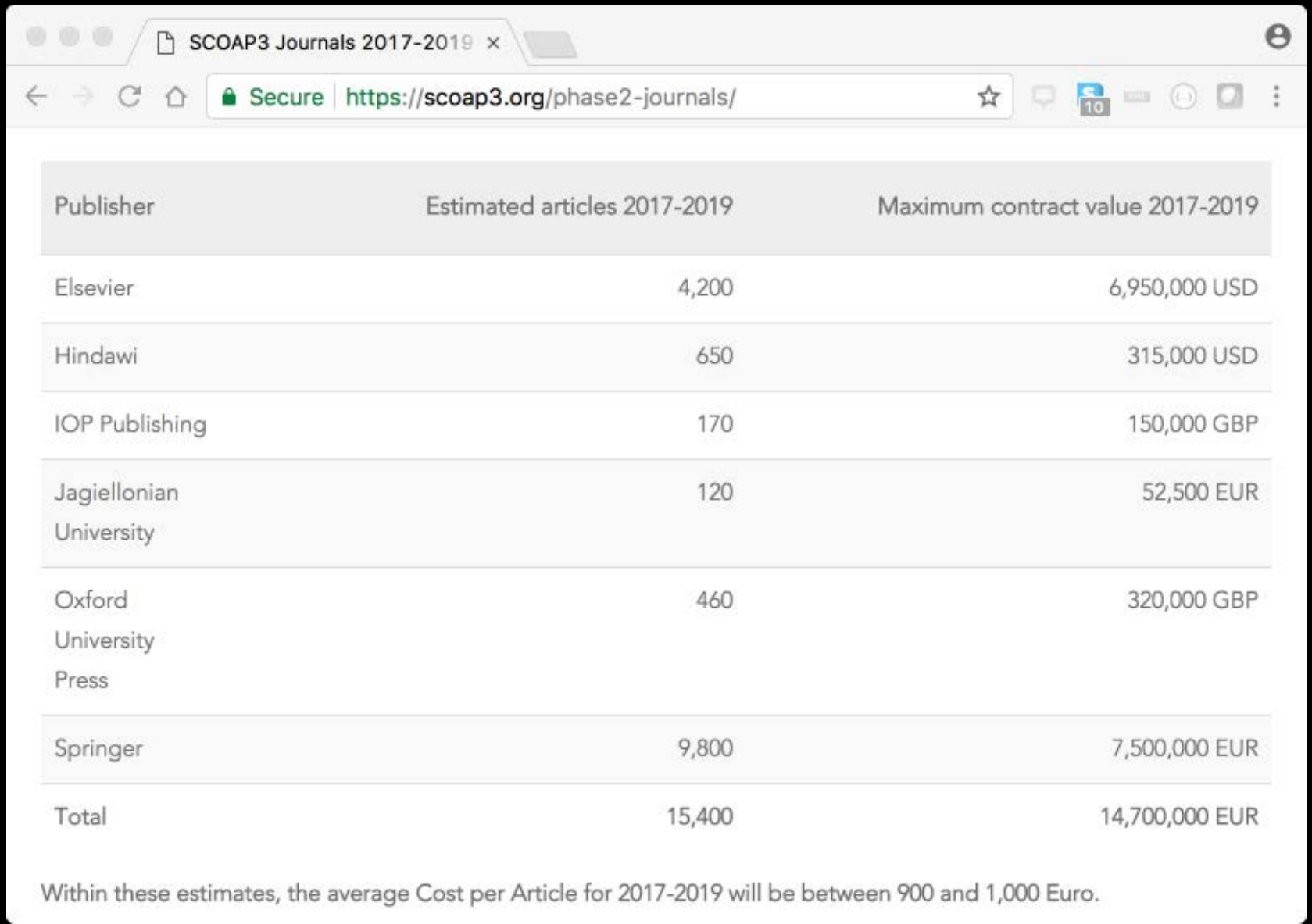
Clicks from SPIRES to DOI links and arXiv in October 2008

All primary (scientific) research outputs should be openly accessible

Why?

Because research will be done more effectively if all shoulders are available to stand on

SCOAP3 contract values

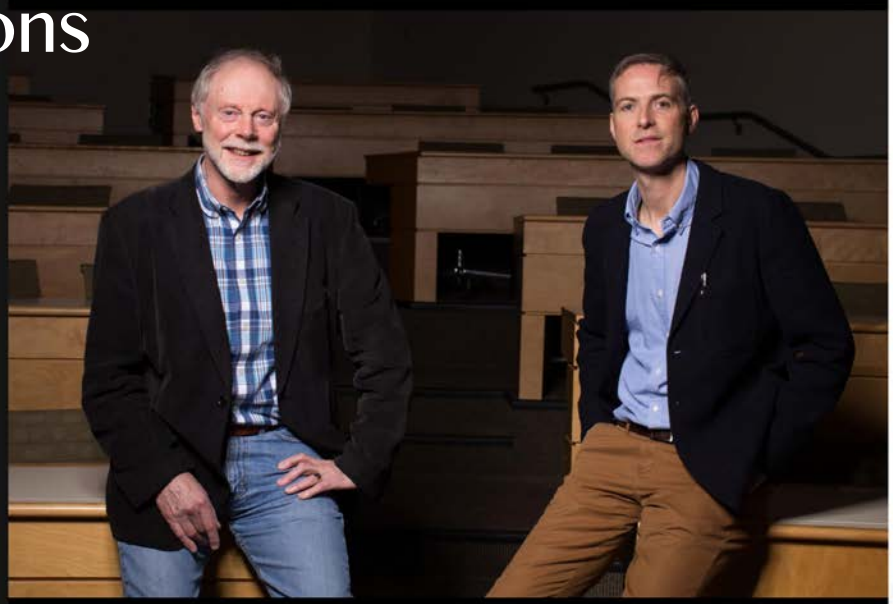


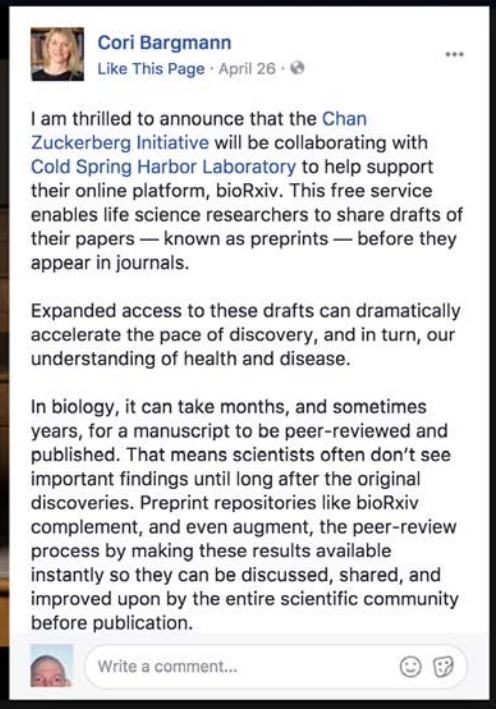
Publisher	Estimated articles 2017-2019	Maximum contract value 2017-2019
Elsevier	4,200	6,950,000 USD
Hindawi	650	315,000 USD
IOP Publishing	170	150,000 GBP
Jagiellonian University	120	52,500 EUR
Oxford University Press	460	320,000 GBP
Springer	9,800	7,500,000 EUR
Total	15,400	14,700,000 EUR

Within these estimates, the average Cost per Article for 2017-2019 will be between 900 and 1,000 Euro.

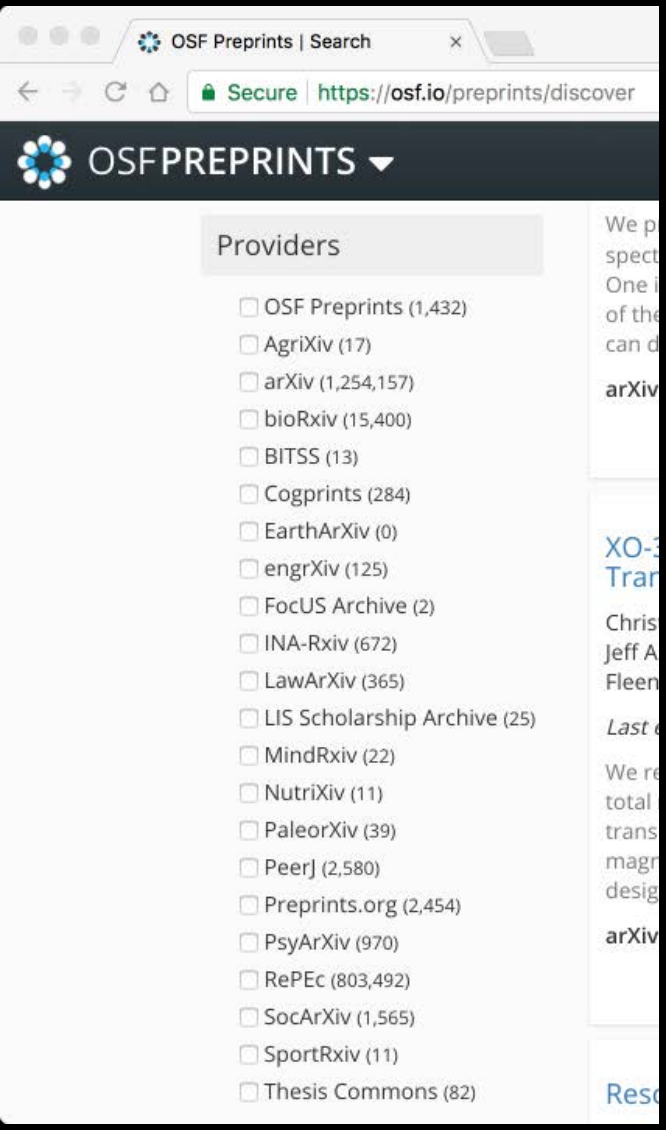
Preprint tipping point?

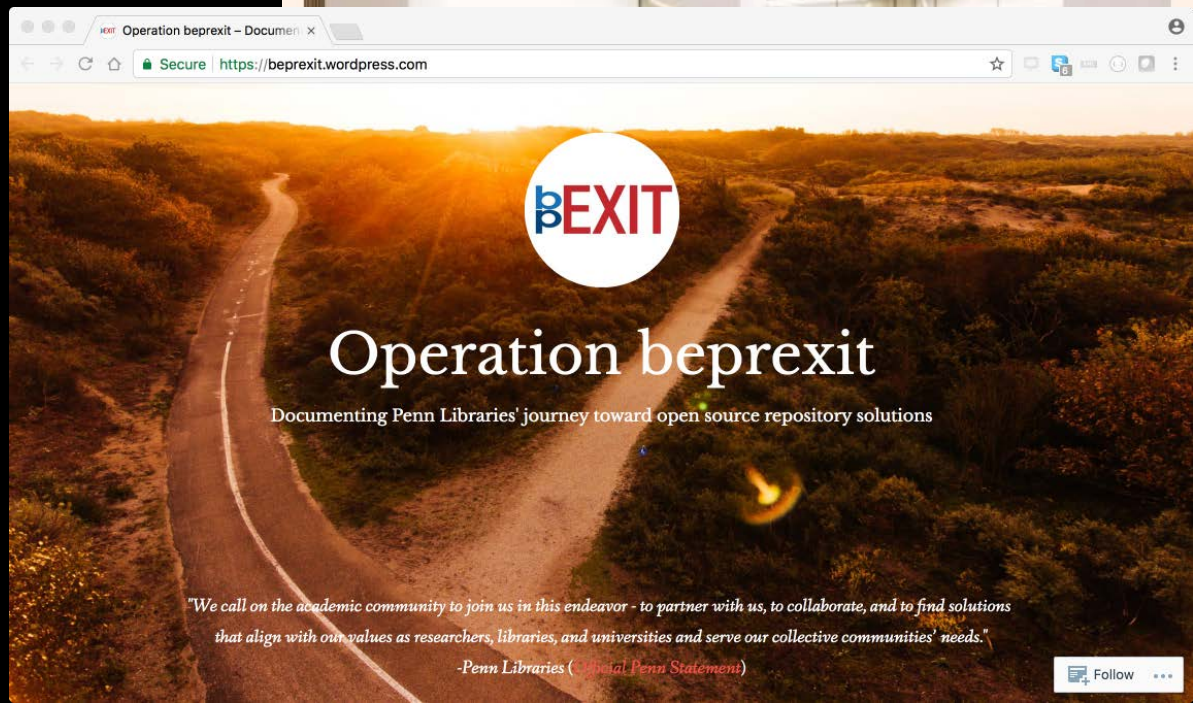
- arXiv “next generation” funding from Sloan and Heising-Simons foundations
- BioRxiv finding from Chen-Zuckerberg
- ASAPbio initiative funded by Sloan, Moore, Arnold and Simons foundations
- …



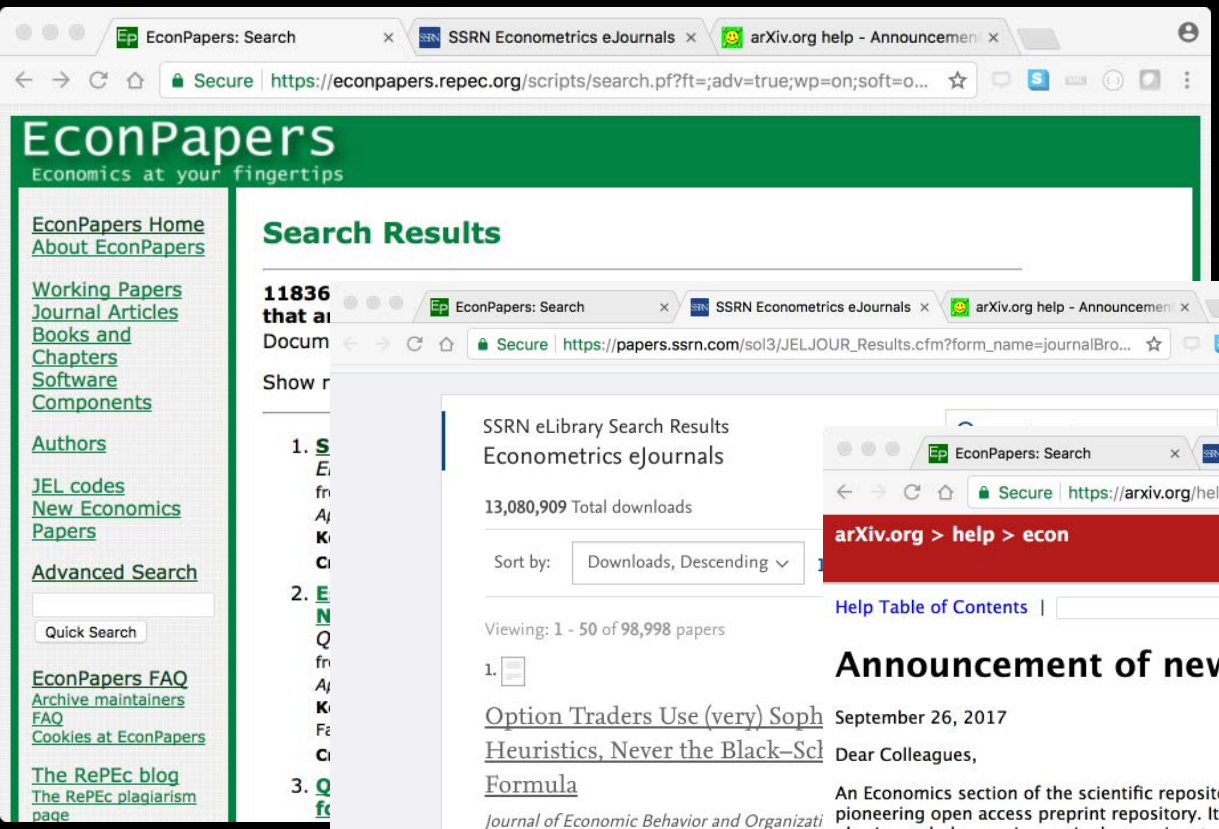


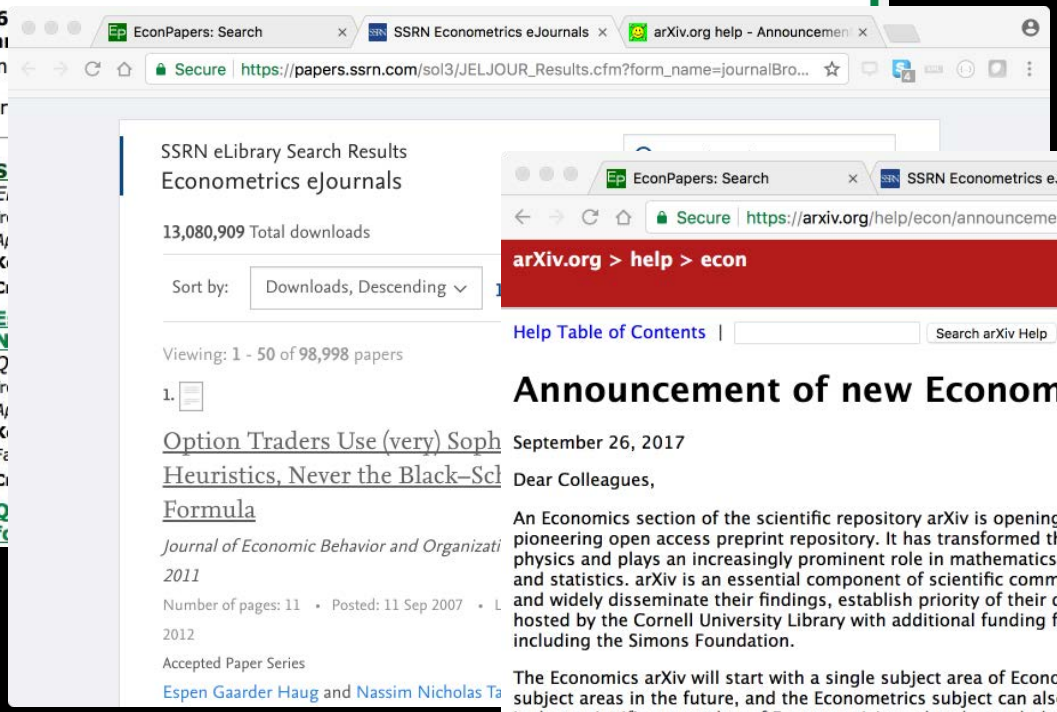
New abcXiv and acquisitions

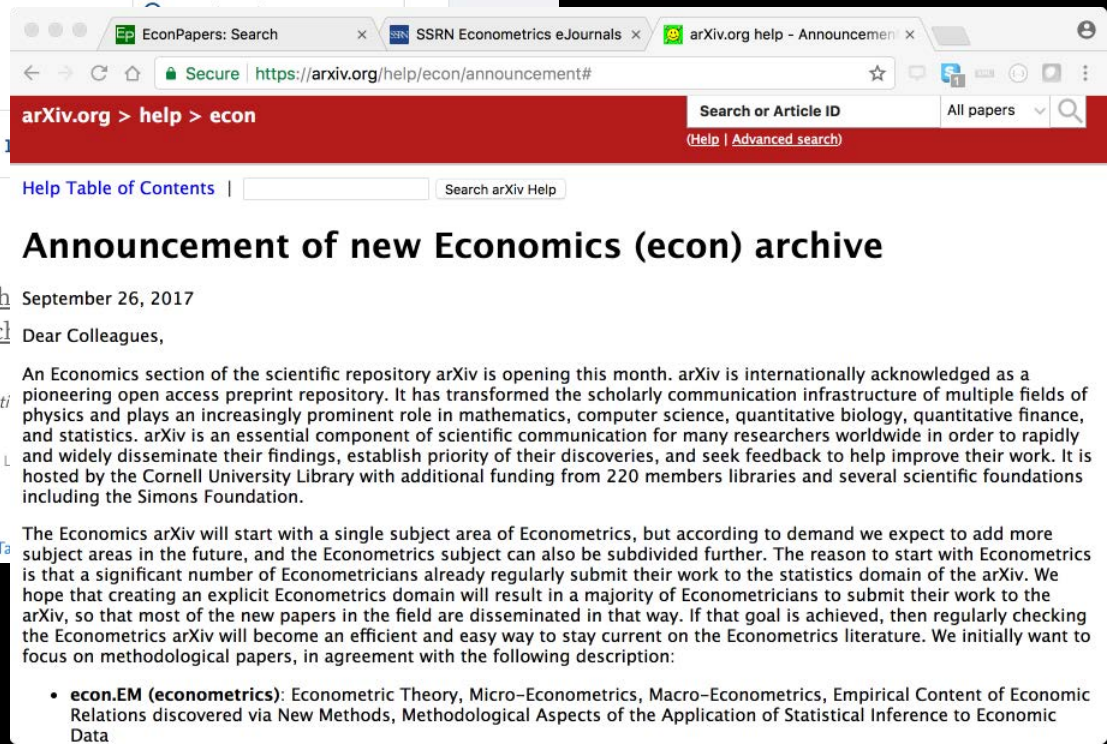




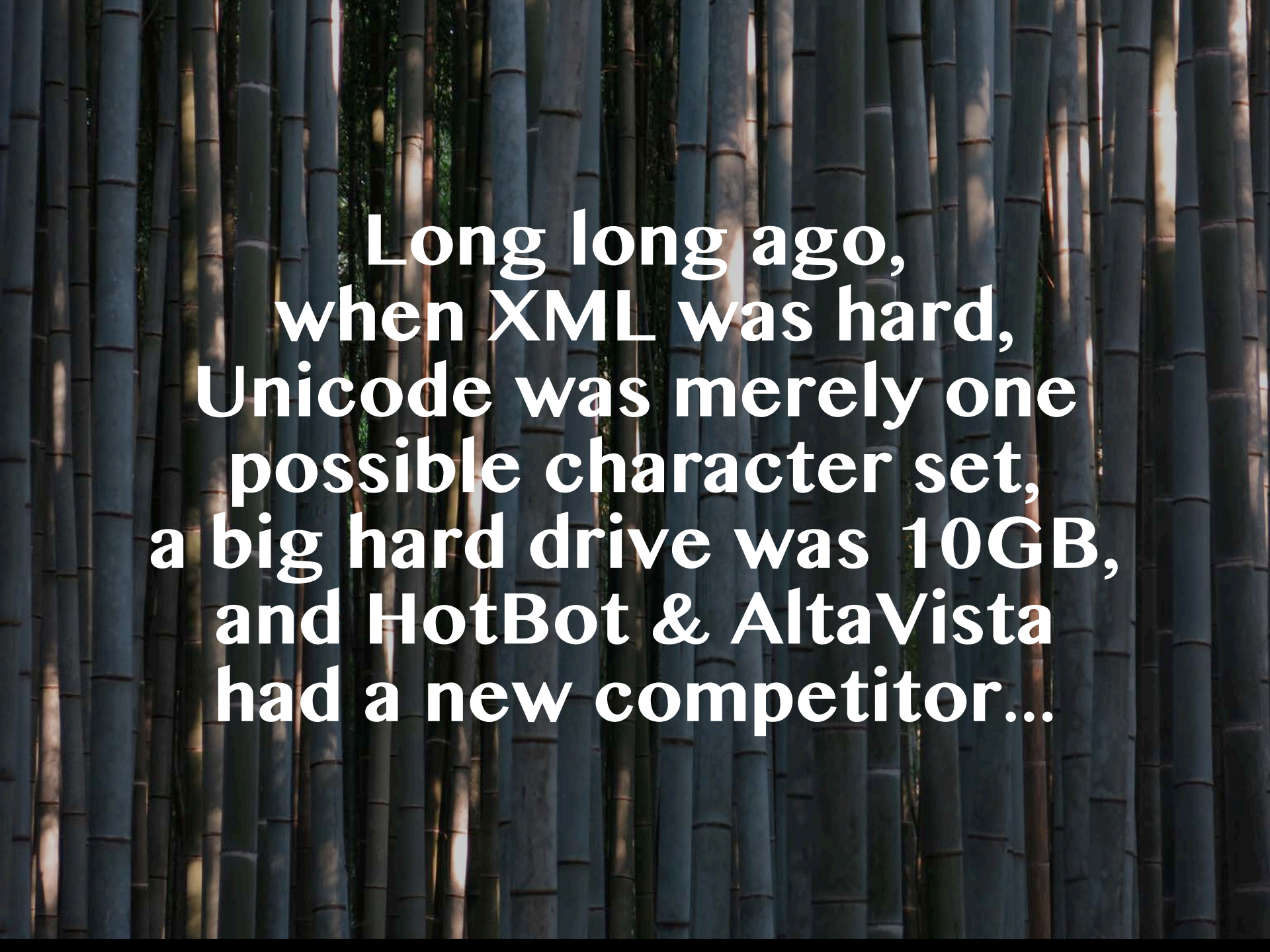
Overlap & competition







Open standards for repository data harvesting

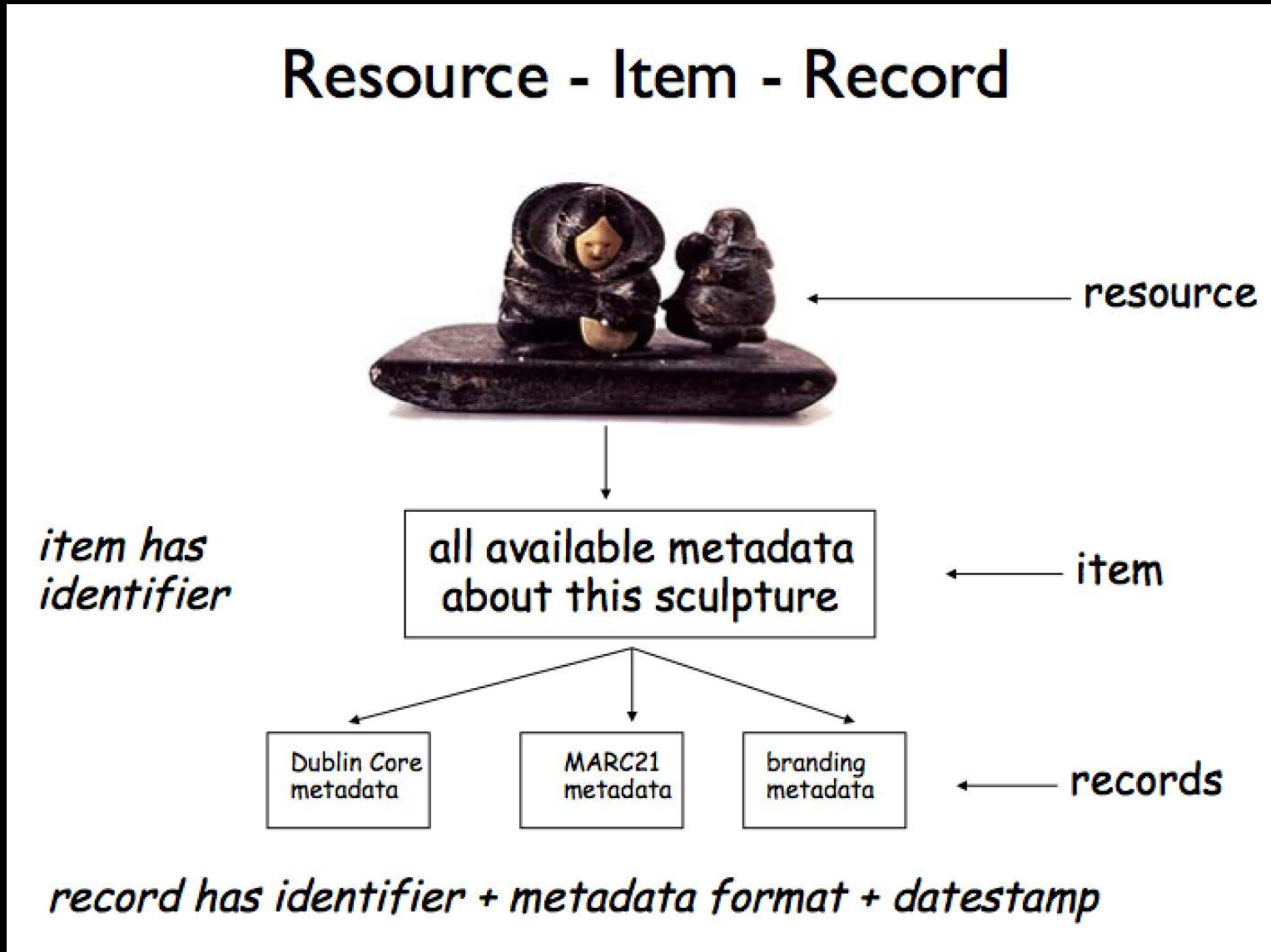

Long long ago,
when XML was hard,
Unicode was merely one
possible character set,
a big hard drive was 10GB,
and HotBot & AltaVista
had a new competitor...

... it was 1999 and the UPS meeting in Santa Fe aimed to

"... identify technologies to stimulate the adoption of the concept of **[Open Access]** ***author self-archived systems in scholarly communication; theorize a framework for the integration of e-print services in the academic document system ..."***

https://www.openarchives.org/meetings/SantaFe1999/ups-invitation-ori.htm

Thus was born OAI-PMH

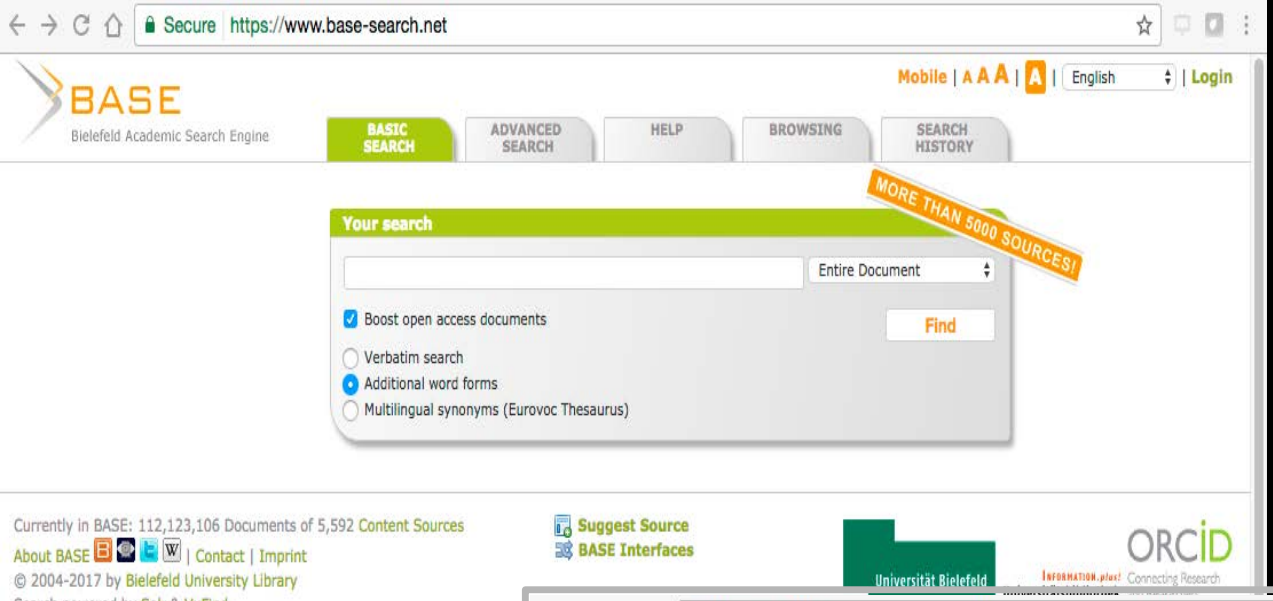


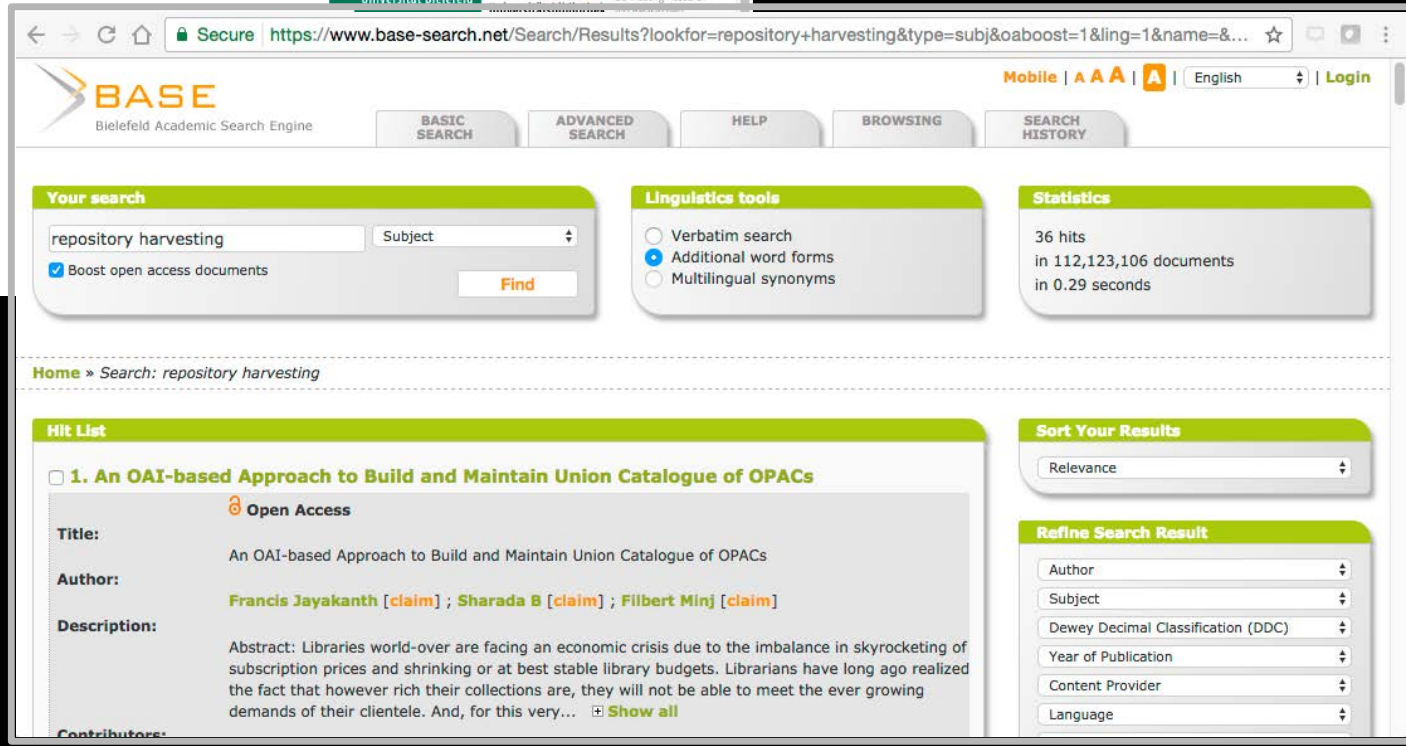
v1.0 2001,
v1.1 2002,
v2.0 2003

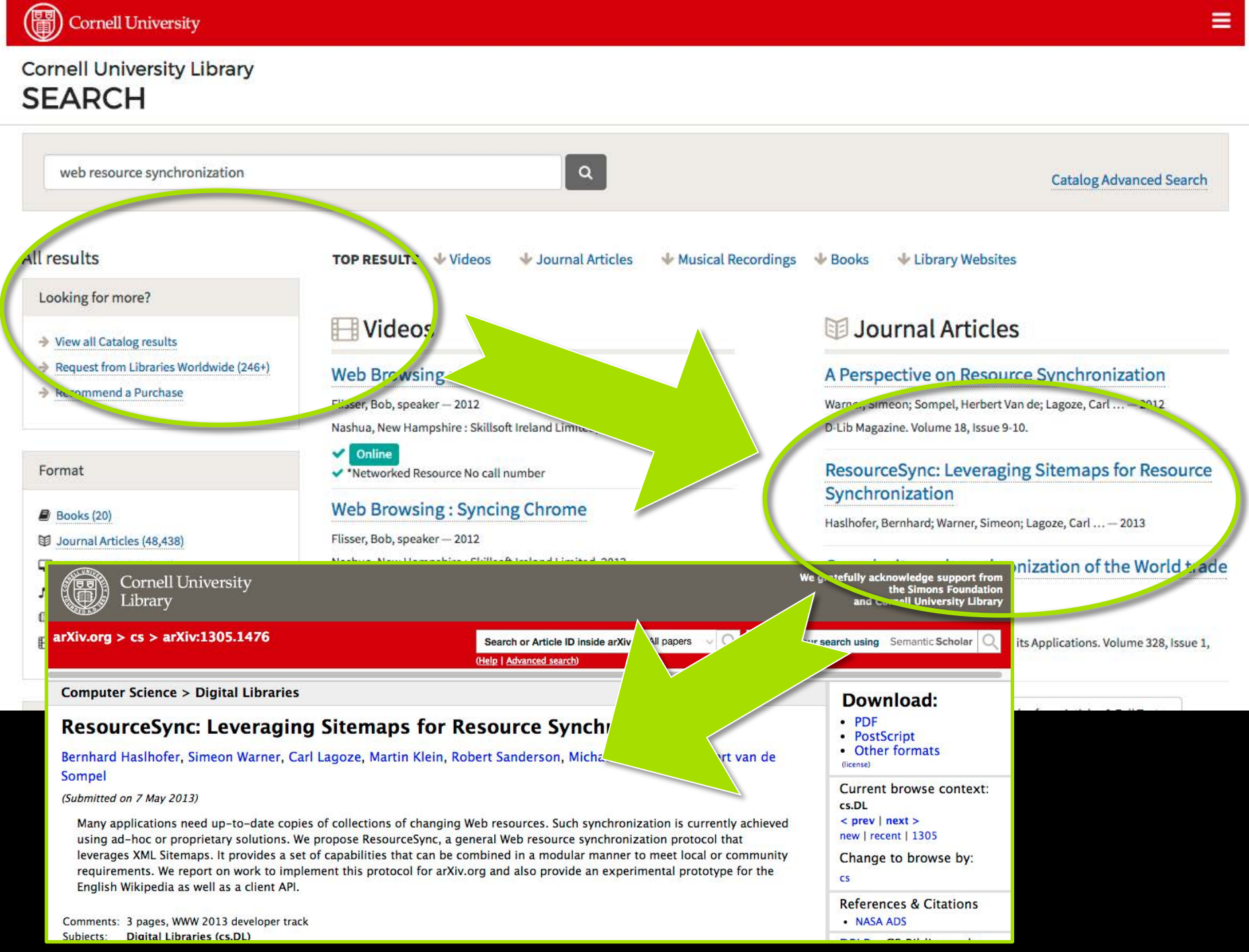
OAI-PMH *was* great!

- It works
- Scales to millions of items
- Easy to implement (good s/w libraries)
- XML, which brought UTF-8 for good multi-language support (hurrah!)
- Widely deployed, stable since 2003 (v2.0)
- Registries & validators
- Community & documentation

BASE harvests >5000 sources >112M documents






Cornell University
Cornell University Library
SEARCH
web resource synchronization
Catalog Advanced Search
All results
TOP RESULTS
Videos
Journal Articles
Musical Recordings
Books
Library Websites
Looking for more?
View all Catalog results
Request from Libraries Worldwide (246+)
Recommend a Purchase
Format
Books (20)
Journal Articles (48,438)
Videos
Web Browsing
Flisser, Bob, speaker — 2012
Nashua, New Hampshire : Skillsoft Ireland Limited
Online
*Networked Resource No call number
Web Browsing : Syncing Chrome
Flisser, Bob, speaker — 2012
Journal Articles
A Perspective on Resource Synchronization
Warner, Simeon; Sompel, Herbert Van de; Lagoze, Carl … — 2012
D-Lib Magazine. Volume 18, Issue 9-10.
ResourceSync: Leveraging Sitemaps for Resource Synchronization
Haslhofer, Bernhard; Warner, Simeon; Lagoze, Carl … — 2013
Cornell University Library
We gratefully acknowledge support from the Simons Foundation and Cornell University Library
arXiv.org > cs > arXiv:1305.1476
Search or Article ID inside arXiv
All papers
Search using Semantic Scholar
(Help | Advanced search)
Computer Science > Digital Libraries
ResourceSync: Leveraging Sitemaps for Resource Synchr
Bernhard Haslhofer, Simeon Warner, Carl Lagoze, Martin Klein, Robert Sanderson, Micha rt van de Sompel
(Submitted on 7 May 2013)
Many applications need up-to-date copies of collections of changing Web resources. Such synchronization is currently achieved using ad-hoc or proprietary solutions. We propose ResourceSync, a general Web resource synchronization protocol that leverages XML Sitemaps. It provides a set of capabilities that can be combined in a modular manner to meet local or community requirements. We report on work to implement this protocol for arXiv.org and also provide an experimental prototype for the English Wikipedia as well as a client API.
Comments: 3 pages, WWW 2013 developer track
Subjects: Digital Libraries (cs.DL)
Download:
PDF
PostScript
Other formats
(license)
Current browse context:
cs.DL
< prev | next >
new | recent | 1305
Change to browse by:
cs
References & Citations
NASA ADS

Technical deficiencies

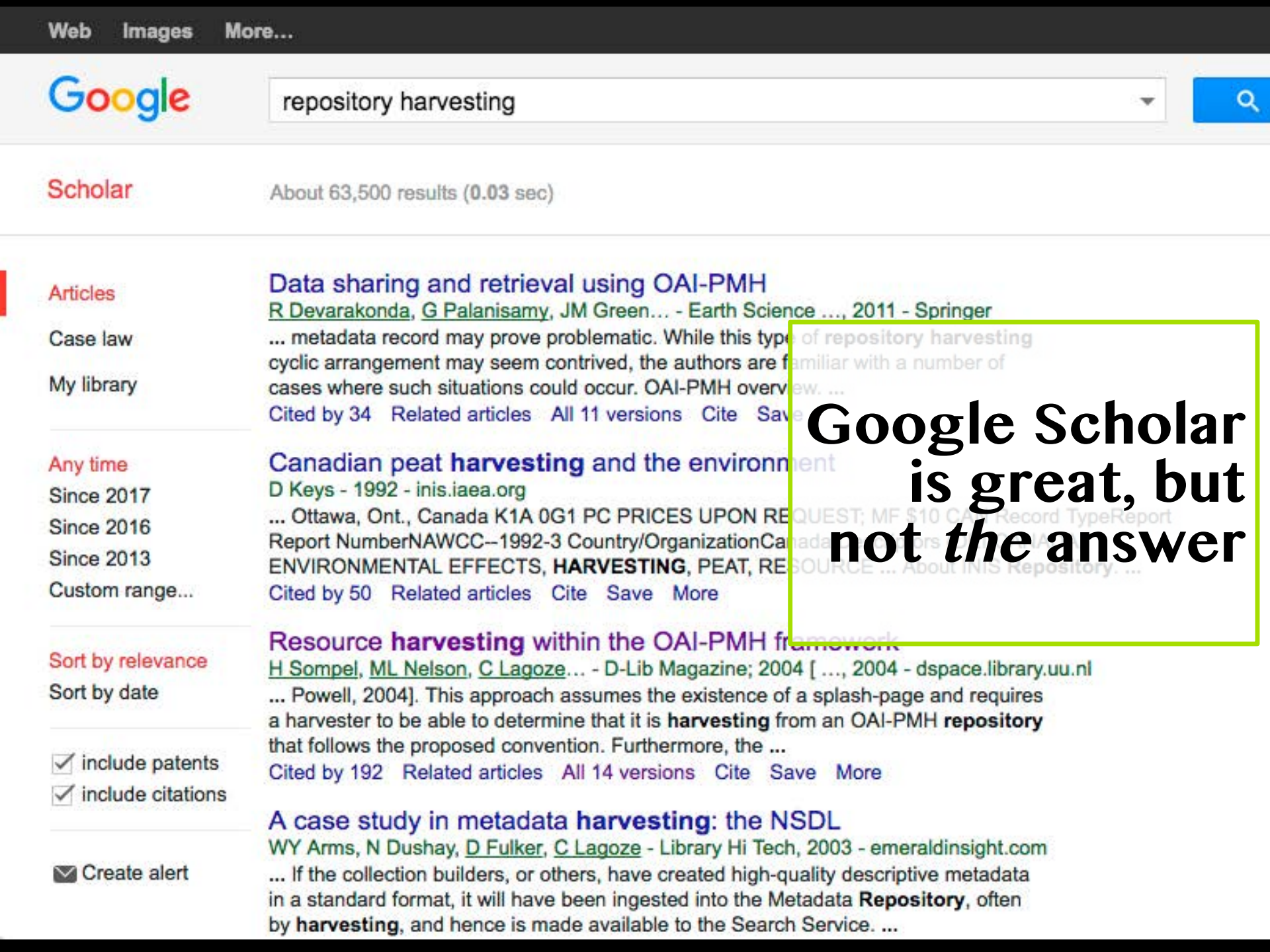
- Not RESTful
- Repository-centric
- XML metadata only
- Metadata is wrapped
- Dynamic set membership bug

"Currently, OAI-PMH is the only behavior that is uniformly exposed by most repositories.

[But]*, its focus on metadata, its pull-based paradigm, and its technological roots that date back to the web of the nineties put it at odds with ... current web technologies."*

COAR Next Generation Repositories
http://comment.coar-repositories.org/2-next-generation-repositories/

Photo by drivethrucafe CC BY-SA
https://www.flickr.com/photos/128758398@N07/15836296662

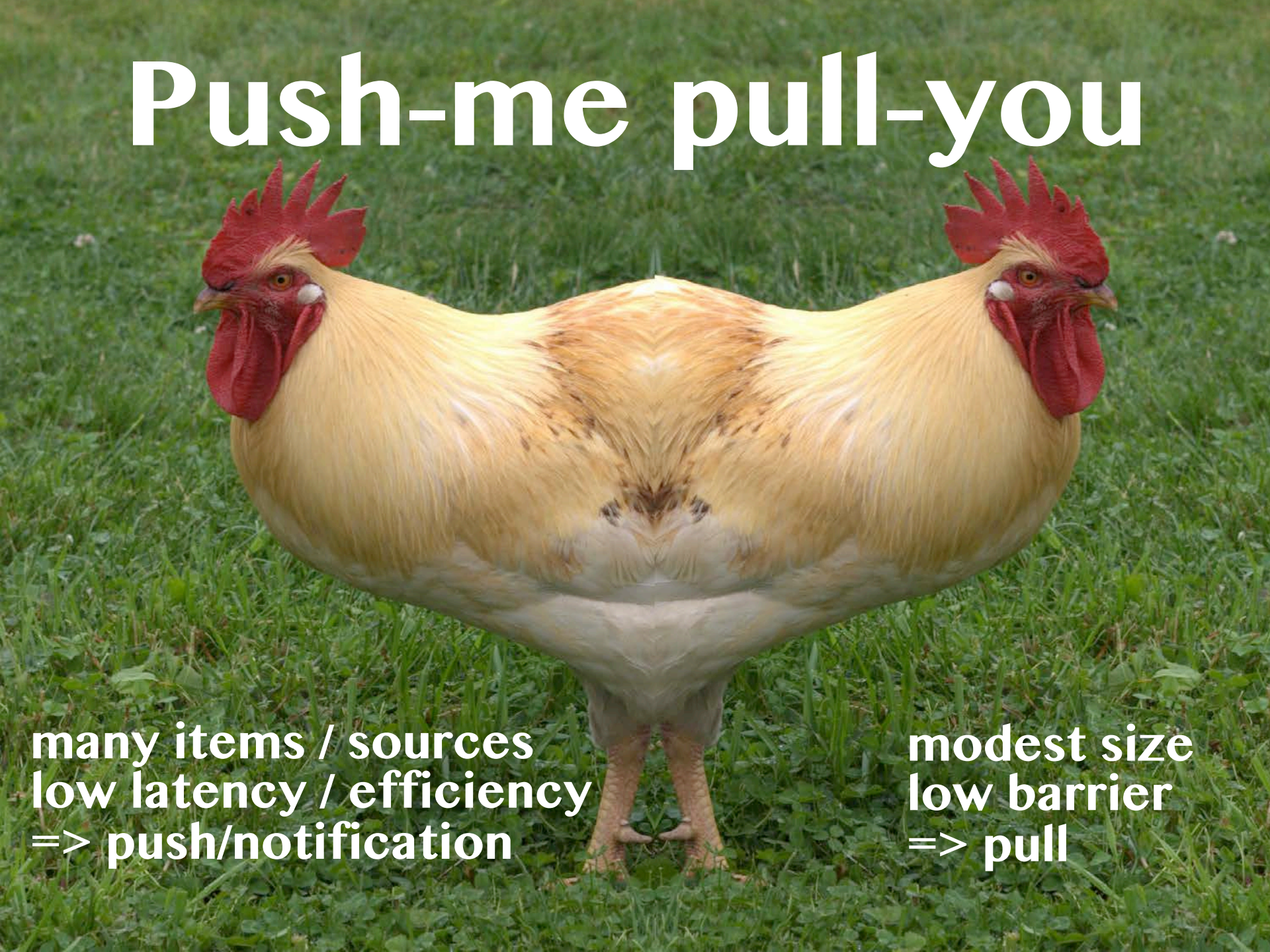

Web Images More...
Google
repository harvesting
Scholar
About 63,500 results (0.03 sec)
Articles
Case law
My library
Any time
Since 2017
Since 2016
Since 2013
Custom range...
Sort by relevance
Sort by date
include patents
include citations
Create alert
Data sharing and retrieval using OAI-PMH
R Devarakonda, G Palanisamy, JM Green... - Earth Science ..., 2011 - Springer
... metadata record may prove problematic. While this type of repository harvesting cyclic arrangement may seem contrived, the authors are familiar with a number of cases where such situations could occur. OAI-PMH overview. ...
Cited by 34 Related articles All 11 versions Cite Save
Canadian peat harvesting and the environment
D Keys - 1992 - inis.iaea.org
... Ottawa, Ont., Canada K1A 0G1 PC PRICES UPON REQUEST; MF $10 CA... Record TypeReport Report NumberNAWCC--1992-3 Country/OrganizationCa... ENVIRONMENTAL EFFECTS, HARVESTING, PEAT, RESOURCE ... About INIS Repository. ...
Cited by 50 Related articles Cite Save More
Resource harvesting within the OAI-PMH framework
H Sompel, ML Nelson, C Lagoze... - D-Lib Magazine; 2004 [..., 2004 - dspace.library.uu.nl
... Powell, 2004]. This approach assumes the existence of a splash-page and requires a harvester to be able to determine that it is harvesting from an OAI-PMH repository that follows the proposed convention. Furthermore, the ...
Cited by 192 Related articles All 14 versions Cite Save More
A case study in metadata harvesting: the NSDL
WY Arms, N Dushay, D Fulker, C Lagoze - Library Hi Tech, 2003 - emeraldinsight.com
... If the collection builders, or others, have created high-quality descriptive metadata in a standard format, it will have been ingested into the Metadata Repository, often by harvesting, and hence is made available to the Search Service. ...
Google Scholar is great, but not the answer

Replacement with no gap

We need a new approach that:

- Meets existing OAI-PMH use cases
- Supports content as well as metadata
- Scales better
- Follows web standards
- Is modern and developer friendly

Push-me pull-you



many items / sources
low latency / efficiency
=> push/notification

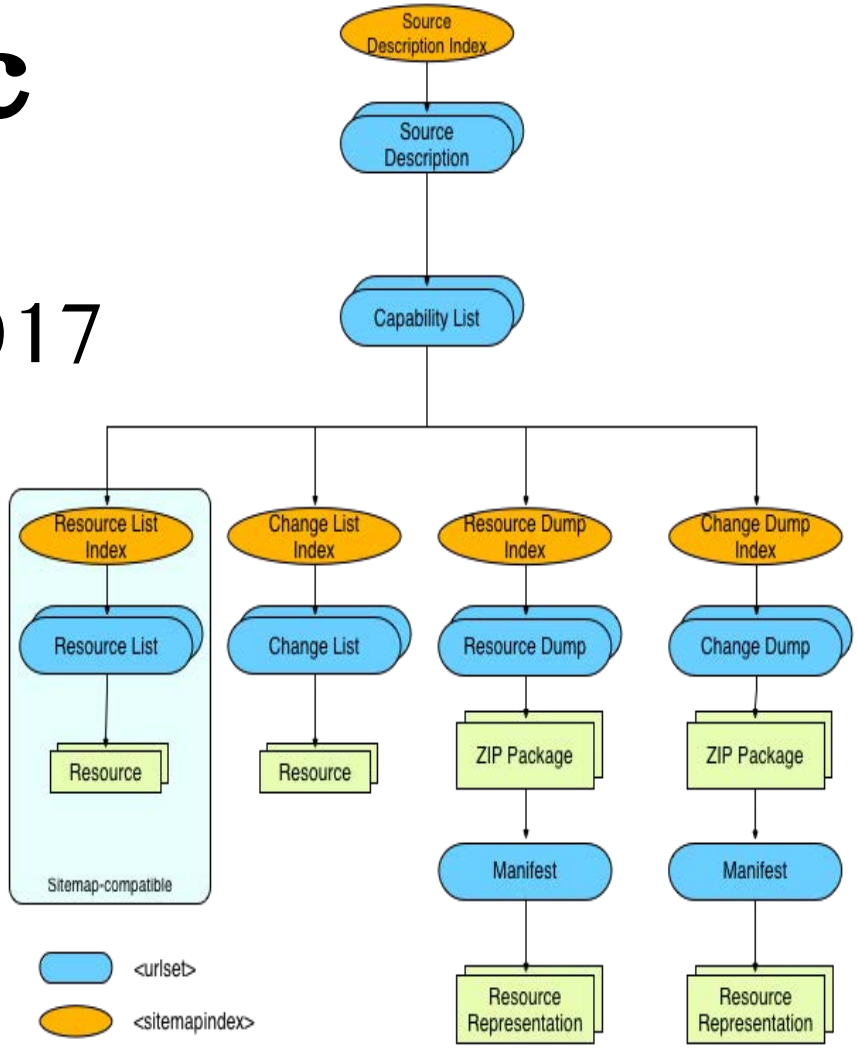
modest size
low barrier
=> pull

ResourceSync

ANSI/NISO Z39.99-2017

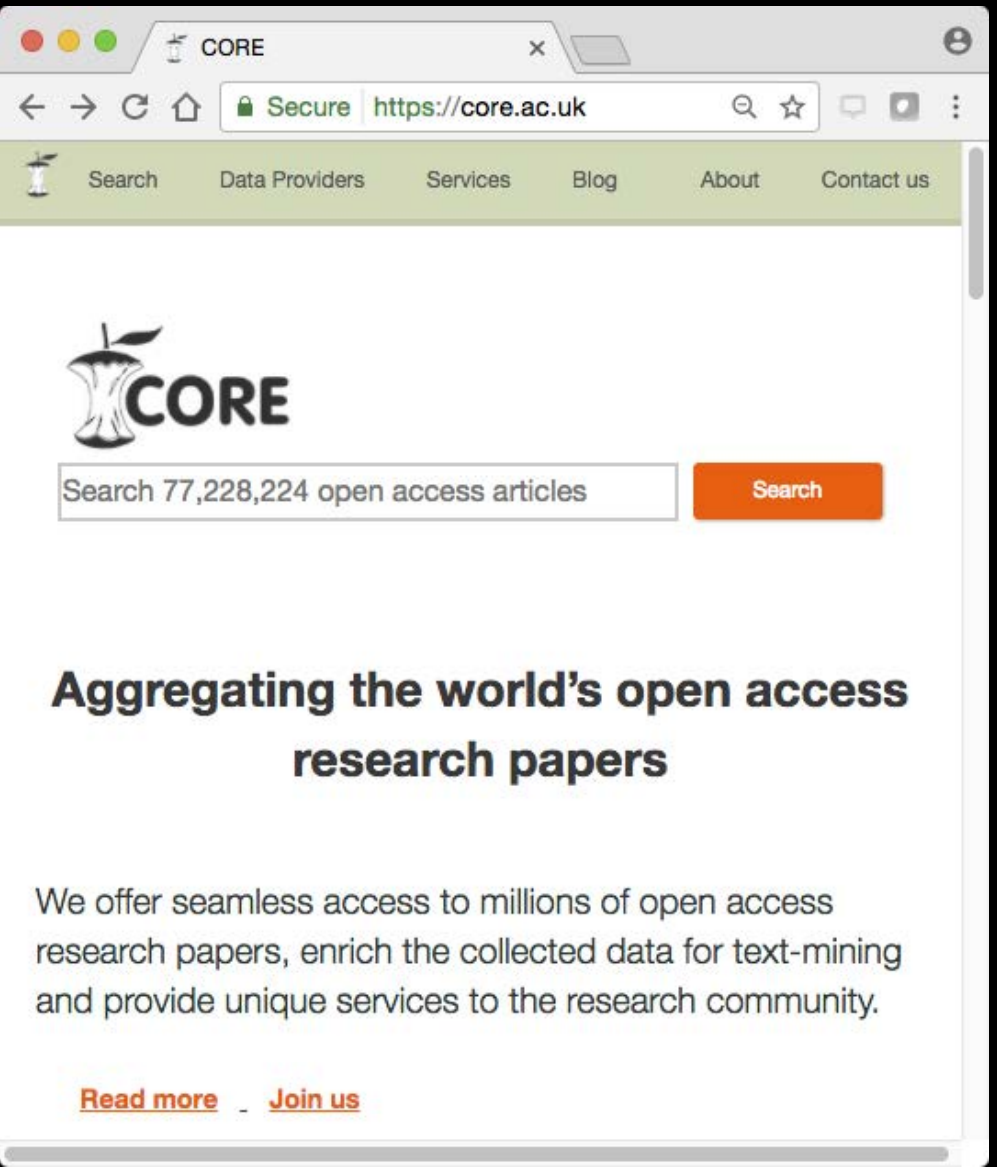
Sitemaps +

- multiple sets
- fixity
- links
- changes only
- dumps



Also supports **Notifications** (push) as optional extension

CORE

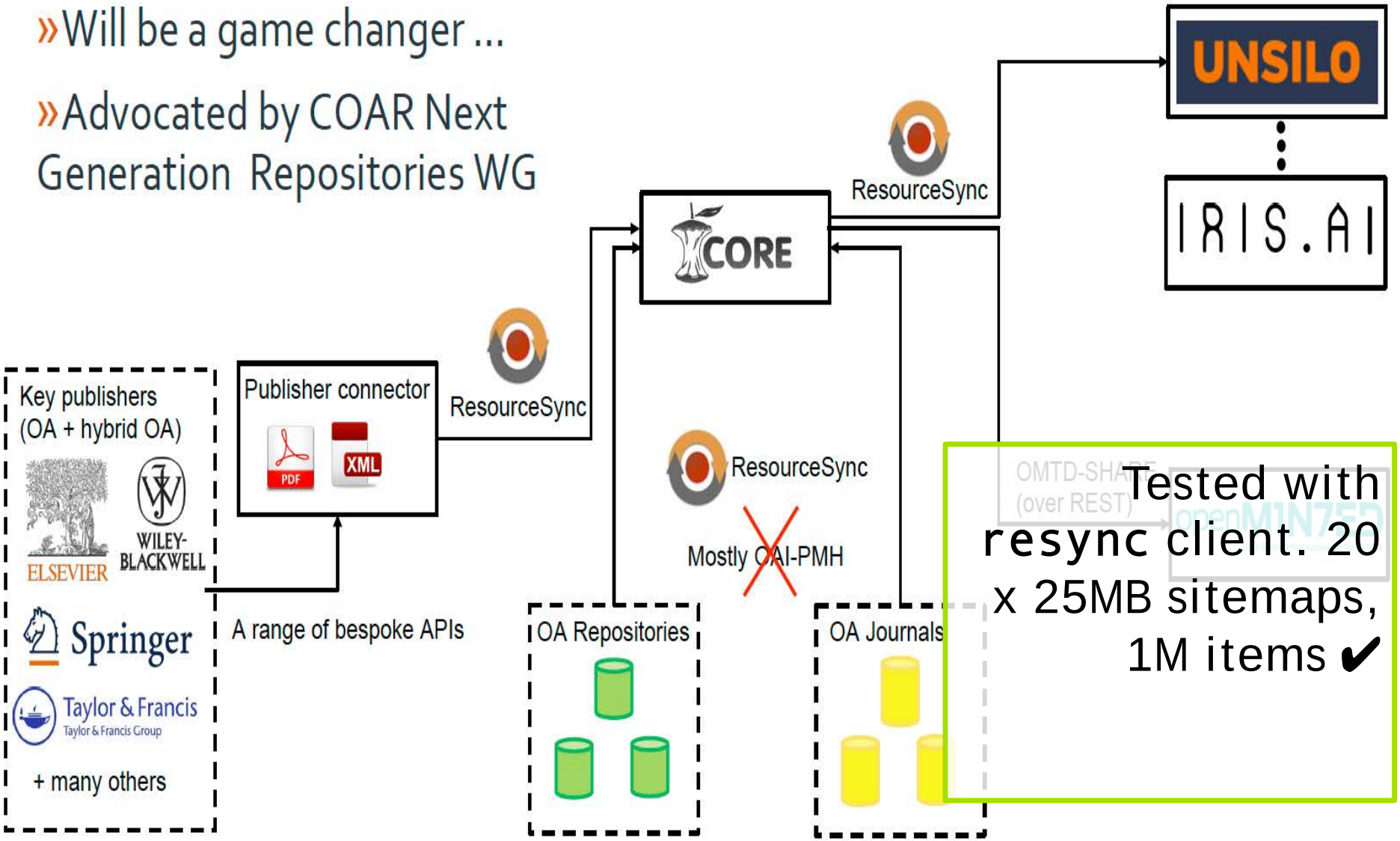


>6000 journals
>2400 repositories
>77M articles
(>6M full text)

metadata +
content

Use Case 3: Replace OAI-PMH with ResourceSync

- Will be a game changer ...
- Advocated by COAR Next Generation Repositories WG



Slide from Petr Knoth / CORE – DPLAfest 2017 presentation -- https://goo.gl/vz3zuJ

Repository Harvesting Conclusions

The repository community should agree on a common new approach to harvesting

ResourceSync was designed to meet

Repository prescription

- Metadata and content should be *web resources*
  - stable URIs, follow web standards, not hidden behind query interfaces
- Support ResourceSync as the primary harvesting interface
  - see e.g. http://hydrainabox.projecthydra.org/2017/06/22/resourcesync.html
  - OAI-PMH as secondary where necessary
- Distinguish and relate metadata and content entries

Person identifiers and ORCID

Some of my person ids

http://orcid.org/0000-0002-7970-7855

http://www.isni.org/isni/0000000351311901

http://www.researcherid.com/rid/E-2423-2011

https://www.scopus.com/authid/detail.uri?authorId=7103063073

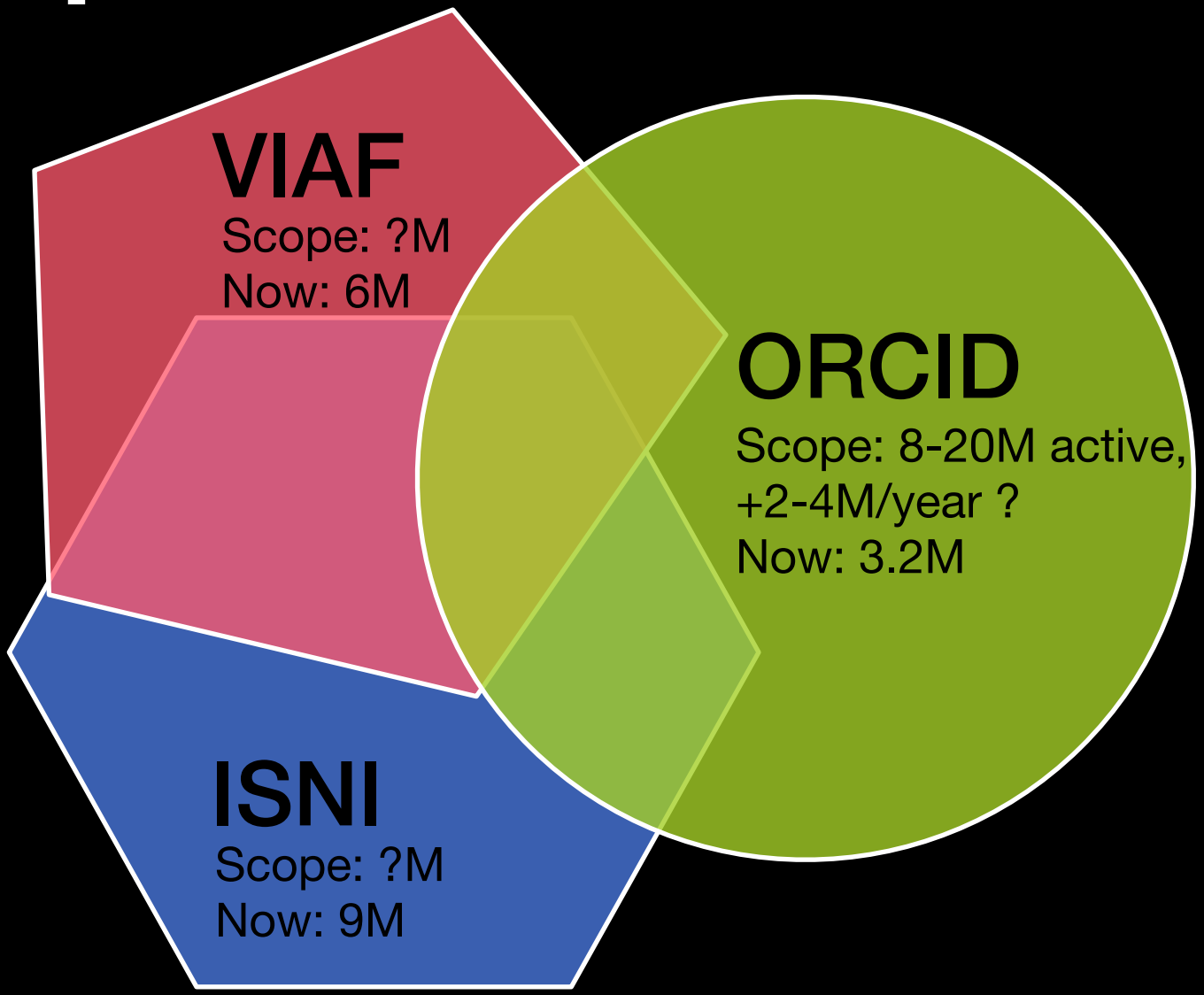
https://arxiv.org/a/warner_s_1

http://vivo.cornell.edu/display/individual24416

https://github.com/zimeon

http://zimeon.com/me

Scopes and scales



Why must ORCID be different?

How many people should have ORCID iDs?

- UNESCO 2013 estimate: 7.8 million researchers
- OECD 2014 estimate: 25.5 million researchers
- Average "active lifetime" 3-6 years (guess)
- Far more than person records in authority systems

How many research and scholarship outputs should be connected to these ORCID iDs?

- ~2 million journal articles published *per year* (https://arxiv.org/abs/1402.4578)
- + >> more if notions of scholarly output extend to data, code, specimens

- "Sort it all out after the fact with manual effort" solution not practical
- Solve with researcher engagement and use in publication workflows

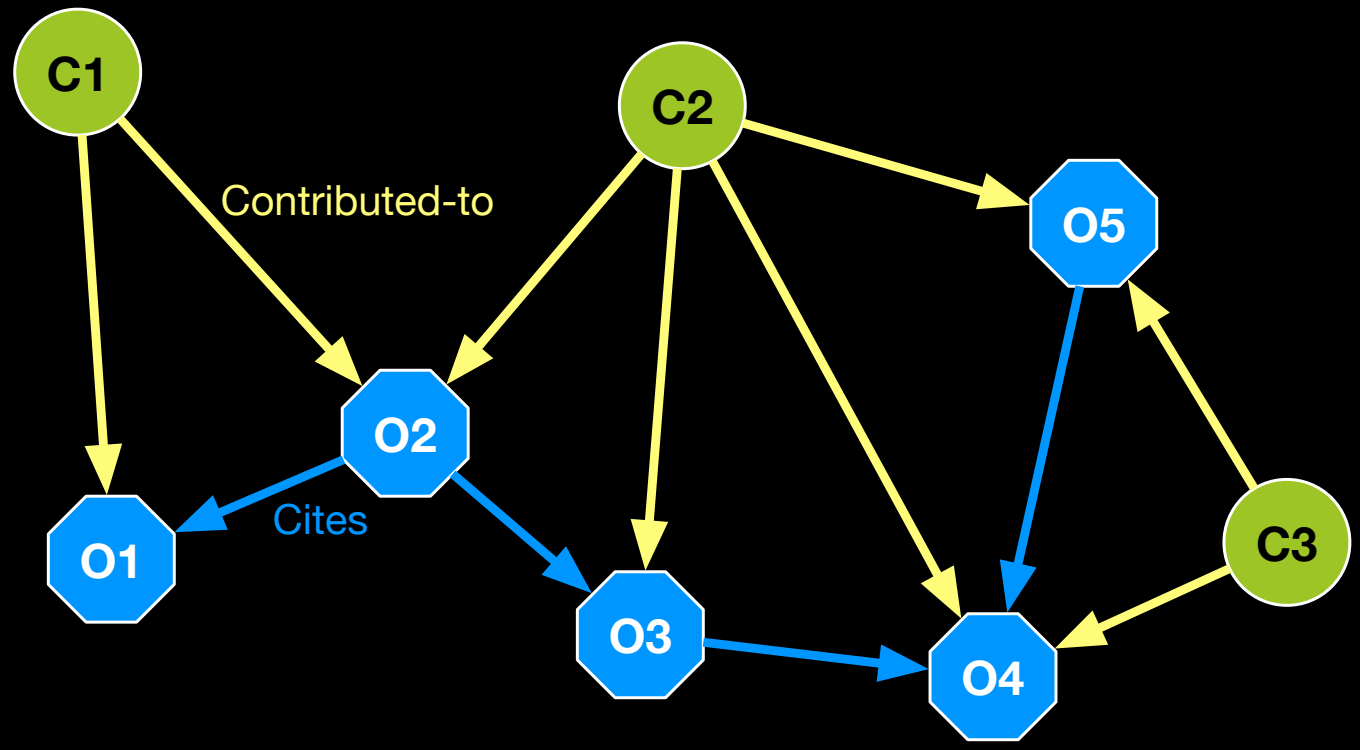
ORCID: Open Researcher and Contributor ID

"ORCID's vision is a world where all who participate in research, scholarship, and innovation are uniquely identified and connected to their contributions across disciplines, borders, and time."

"ORCID provides an identifier for individuals to use with their name as they engage in research, scholarship, and innovation activities. We provide open tools that enable transparent and trustworthy connections between researchers, their contributions, and affiliations. We provide this service to help people find information and to simplify reporting and analysis." (https://orcid.org/)

- Research and scholarship focus
- Expect use by individuals identified in workflows

Contributor-Output graph

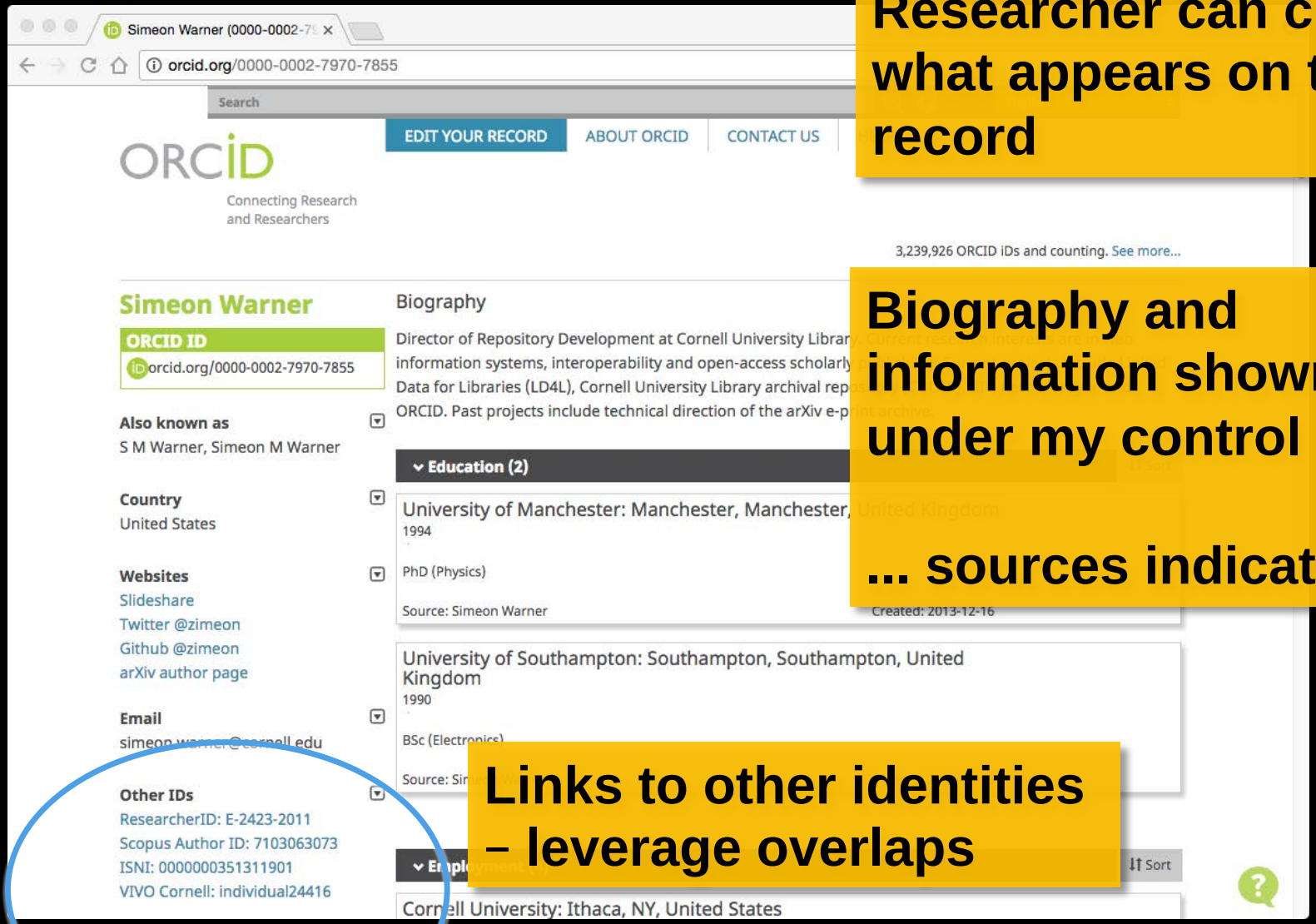


Generalize:

- many contributor roles
- expand “cites” to include other notions of derivation
- ++ add organization nodes for affiliation/funding/ etc. (and time dependence)

For full benefit ORCID needs most researchers to willingly use their ORCID iD.

Researcher control



Researcher can choose what appears on their record

Biography and information shown under my control

... sources indicated

Links to other identities – leverage overlaps

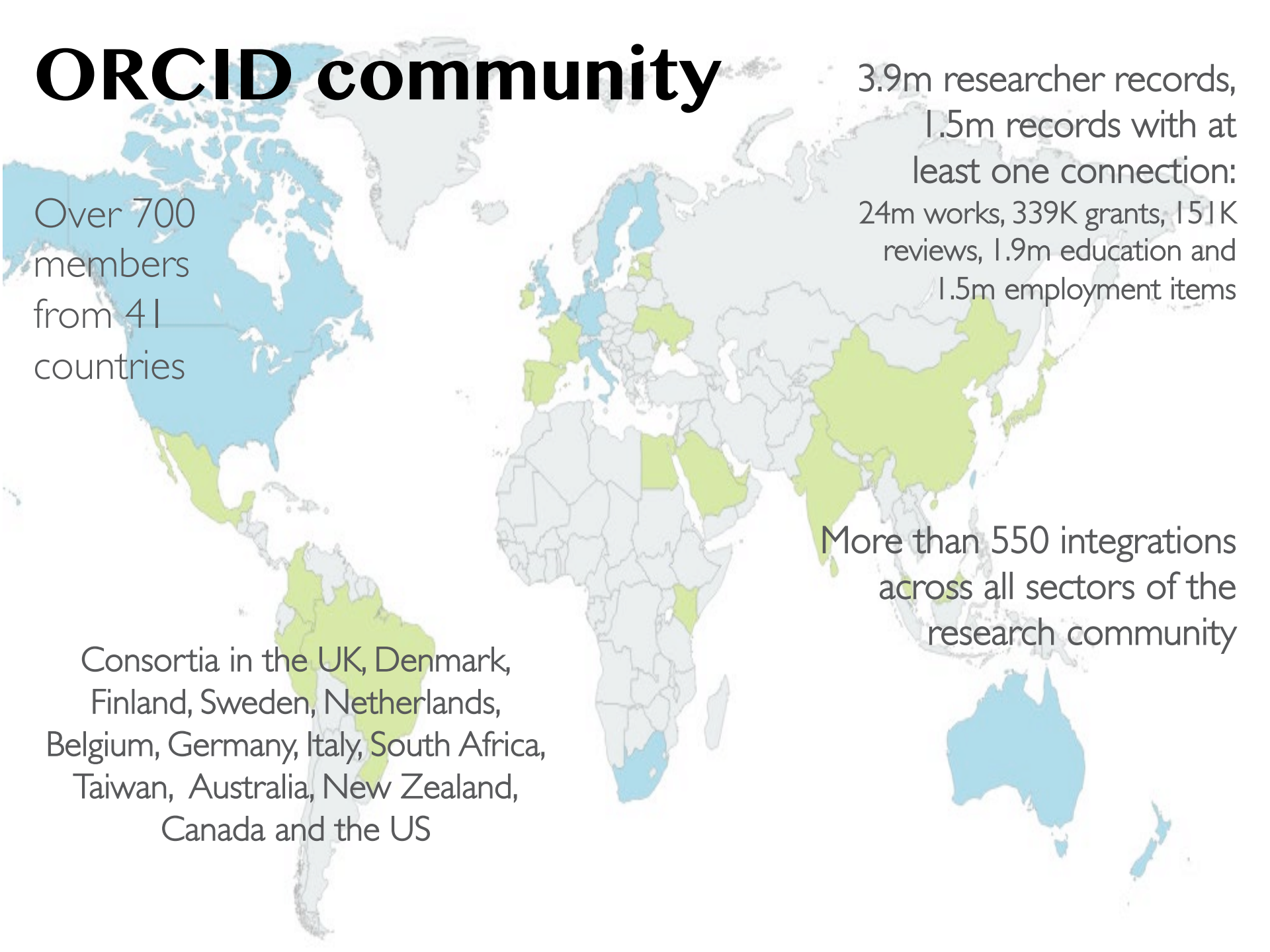
ORCID iD use

- 7000 journals use ORCID iDs, over 1500 of which require use by corresponding authors
- Researcher support from surveys:
  - In 2017 85.9% of respondents now believe requiring the use of ORCID iDs is beneficial to the global research community, compared with 72.2% of 2015 respondents
  - In 2017 83.1% of respondents strongly agree/agree that ORCID is “essential”, compared with 48.8% in 2015.

ORCID community

Over 700 members from 41 countries

3.9m researcher records, 1.5m records with at least one connection:
24m works, 339K grants, 151K reviews, 1.9m education and 1.5m employment items



More than 550 integrations across all sectors of the research community

Consortia in the UK, Denmark, Finland, Sweden, Netherlands, Belgium, Germany, Italy, South Africa, Taiwan, Australia, New Zealand, Canada and the US

Finland's national ORCID cons ×
Secure | https://researcheridentifier.fi/finlands-national-orcid-consortium/

FINLAND'S NATIONAL ORCID CONSORTIUM

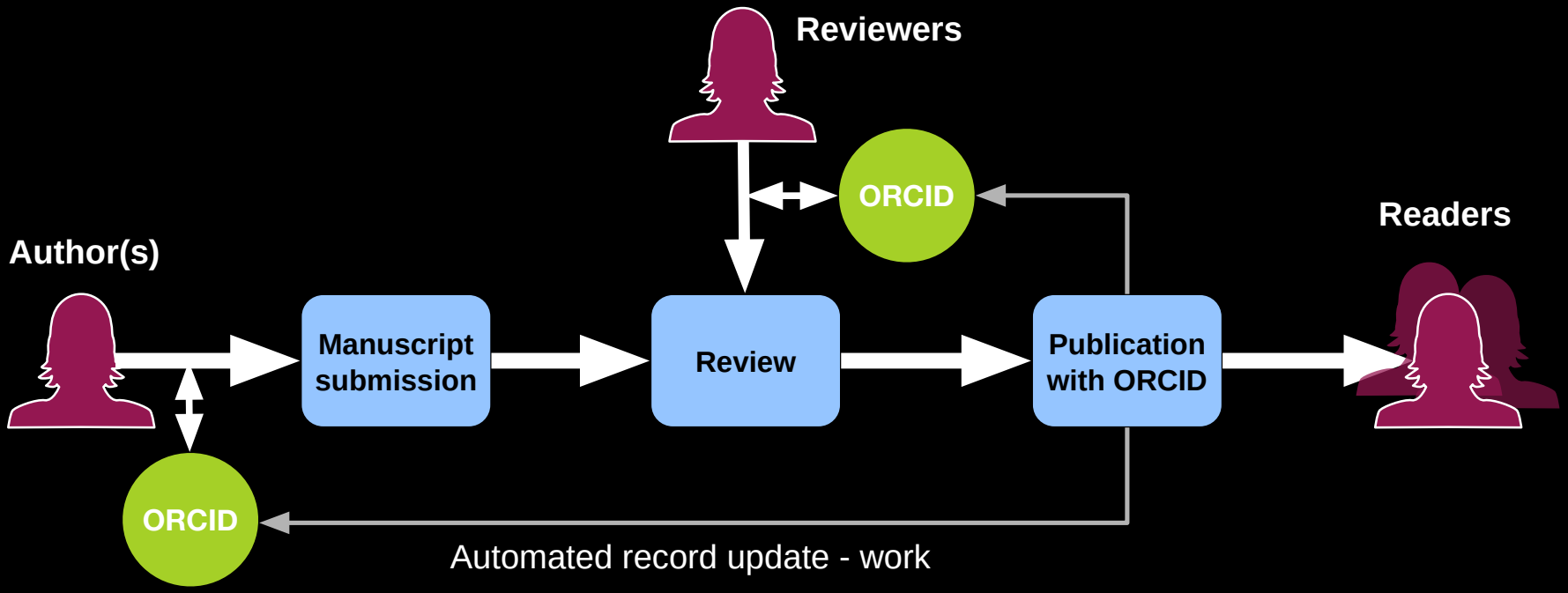
Finland's national ORCID consortium was established in May 2016 and launched on 1 June 2016. Thirteen Finnish research organisations have initially joined the consortium, along with CSC – IT Center for Science, which has been commissioned by the Ministry of Education and Culture to act as the consortium's coordinator. Consortium membership is open to all interested Finnish organisations.

CONSORTIUM MEMBERS:

- CSC – IT Center for Science Ltd
- Hanken School of Economics
- University of Helsinki
- Helsinki and Uusimaa Hospital District
- Finnish Meteorological Institute
- University of Jyväskylä
- Pirkanmaa Hospital District
- Hospital District of Northern Savo
- University of Arts Helsinki
- Tampere University of Technology
- The National Institute for Health and Welfare (THL)
- University of Turku
- University of Vaasa
- Hospital District of Southwest Finland

ORCID Stakeholders, Actions and Benefits

Journal article round trip



ORCID iDs are intended to be integrated into research and publication workflows, and become embedded in metadata. Thus ORCID iDs associated with works when published

- Ambiguity avoidance rather than disambiguation!

Linked Open Data

Not (quite) the semantic web

"it is clearly a good idea, and some very nice demonstrations exist, but it has not yet changed the world"

[out of context quote from "The Semantic Web" Berners-Lee, Hendler and Lassila, Scientific American, May 17, 2001]

Linked Data

- A practical "sematic web lite"
- Narrower focus

("RDF standards" such as ontologies, SPARQL, etc. are the gateway to a more complete semantic web.)

https://www.w3.org/DesignIssues/LinkedData.html

Why replace MARC with Linked Data formats?

1. MARC is inadequate

MARC continues to meet many needs, but there are several areas of stress:

- Translation of record, not descriptions of appropriate entities
- Use of text when we want data
- Limited extensibility
- Imprecise URI references (record or RWO?)
- ...

2. Use identifiers not names

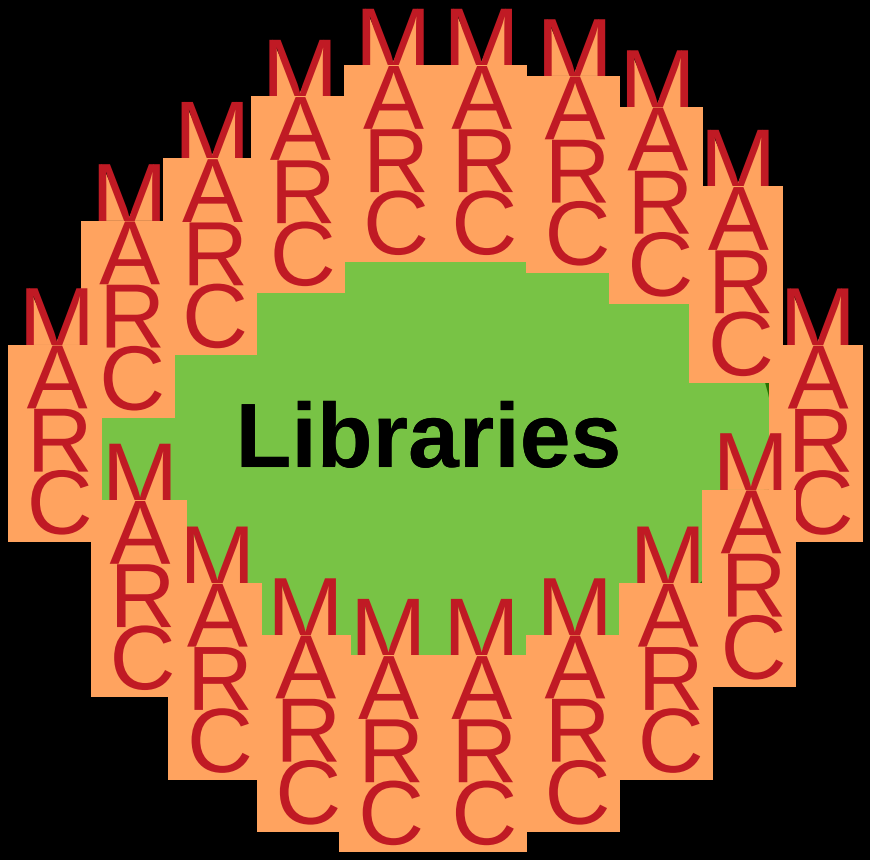
Identifiers provide necessary layer of indirection that authorized names do not:

- Identifiers more easily stable
  - e.g. no change from “Banks, Iain, 1953-” to “Banks, Iain, 1953-2013”
- Exact matching
- URIs make the web work well
- Does not replace authority ideas, just makes them work better

3. Connect to the web

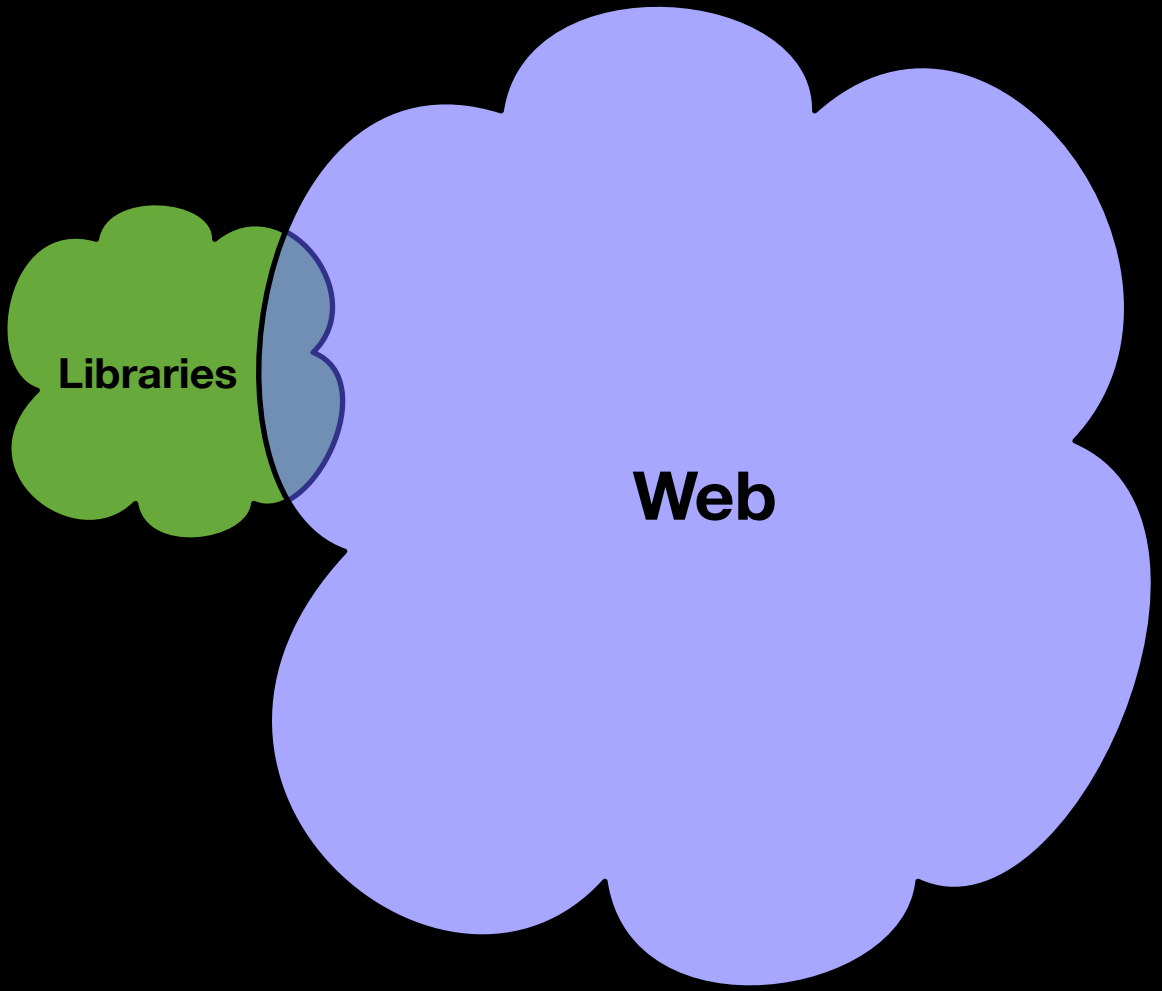
"Fortress MARC" protects and isolates libraries from the web

- Little reuse of our data
- Can't use standard tools
- Difficult to generalize



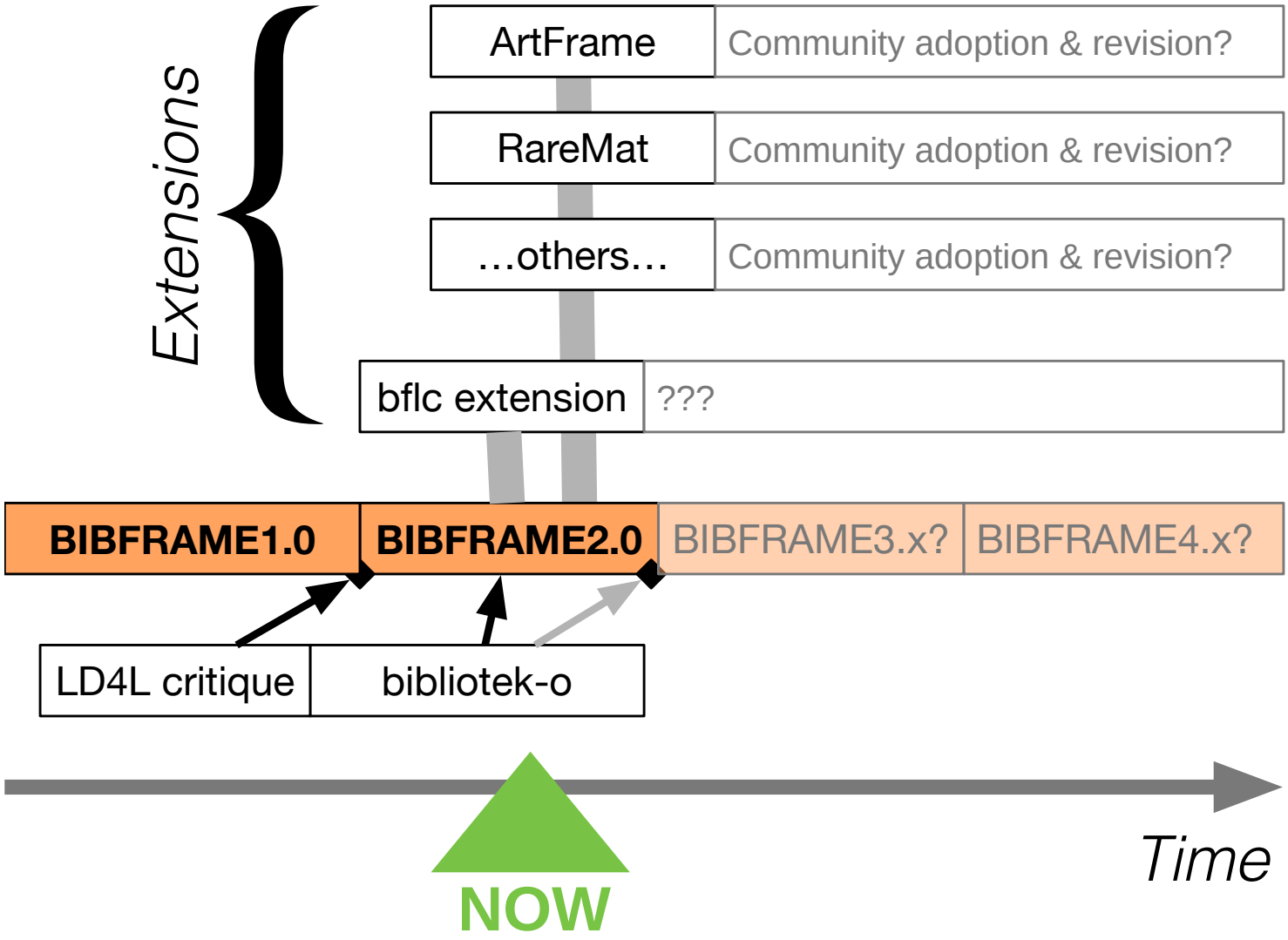
The web is big ...

... most of our users spend most of their time there



[not to scale]

BIBFRAME & related ontologies



LD4L & LD4L Labs

https://ld4l.org/ld4l-labs/

Cornell, Harvard, Stanford, Iowa; 2014-2016

- Conversion of MARC -> BIBFRAME at scale (~30M records, ~3billion triples)
- Blacklight-based search over combined catalogs
- Ontology work around “LD4L ontology” which provided significant input for BIBFRAME2.0
- Support use of linked data authorities in the Hydra stack via Questioning Authority gem

2016-2018

- bibliotek-o ontology
- Data conversion MARC & non-MARC to LD
- VitroLib editor
- Authority infrastructure and UI refinement including context

LD4P – … for Production

https://ld4l.org/ld4p

Columbia, Cornell, Harvard, LC, Princeton, Stanford – 2016-2018

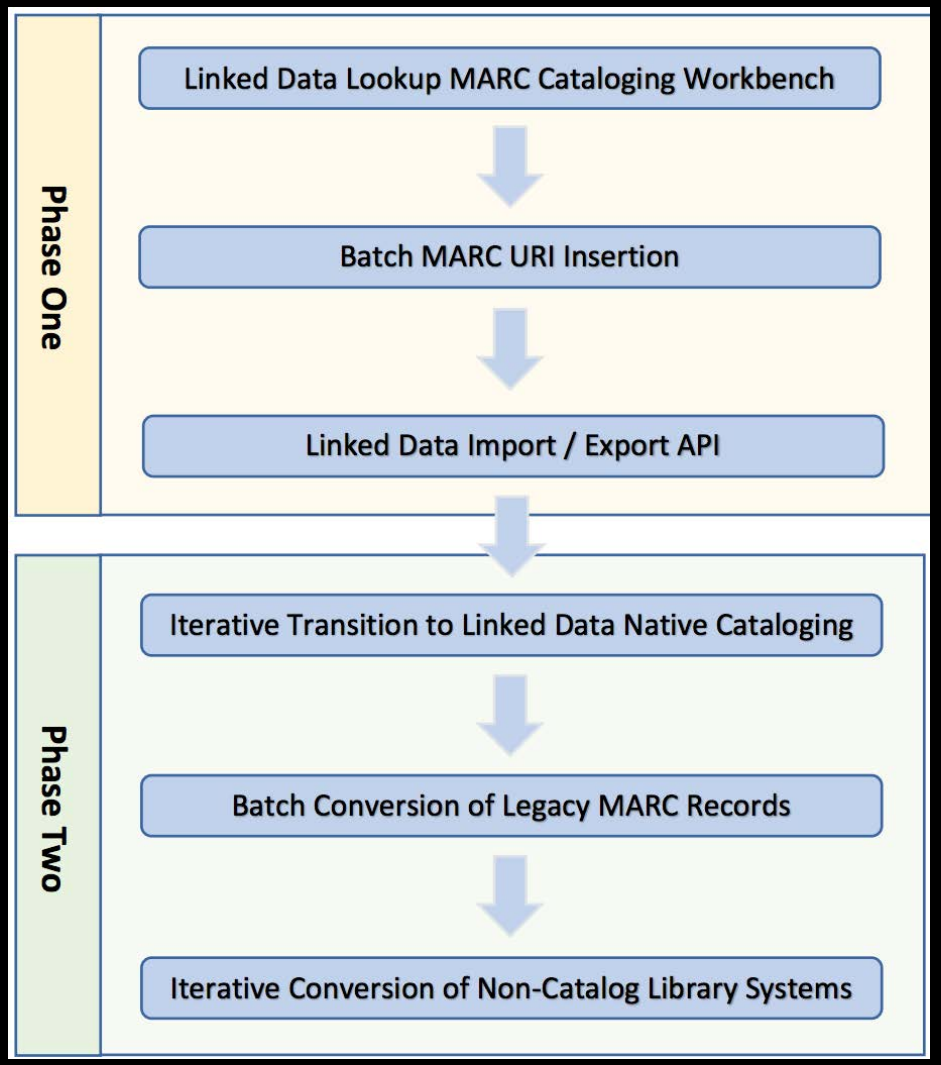
- Develop extension ontologies for BIBFRAME2.0/bibliotek-o (ArtFrame, Cartographic, Moving Image, Performed Music, & Rare Materials)
- Pilot transition of technical services workflows to a linked data environment
  - copy cataloging
  - original cataloging

("production" in LD4P means creation of catalog records, not production-ready)

BIBFLOW (UCDavis, 2014-2016)

Conservative suggestion:

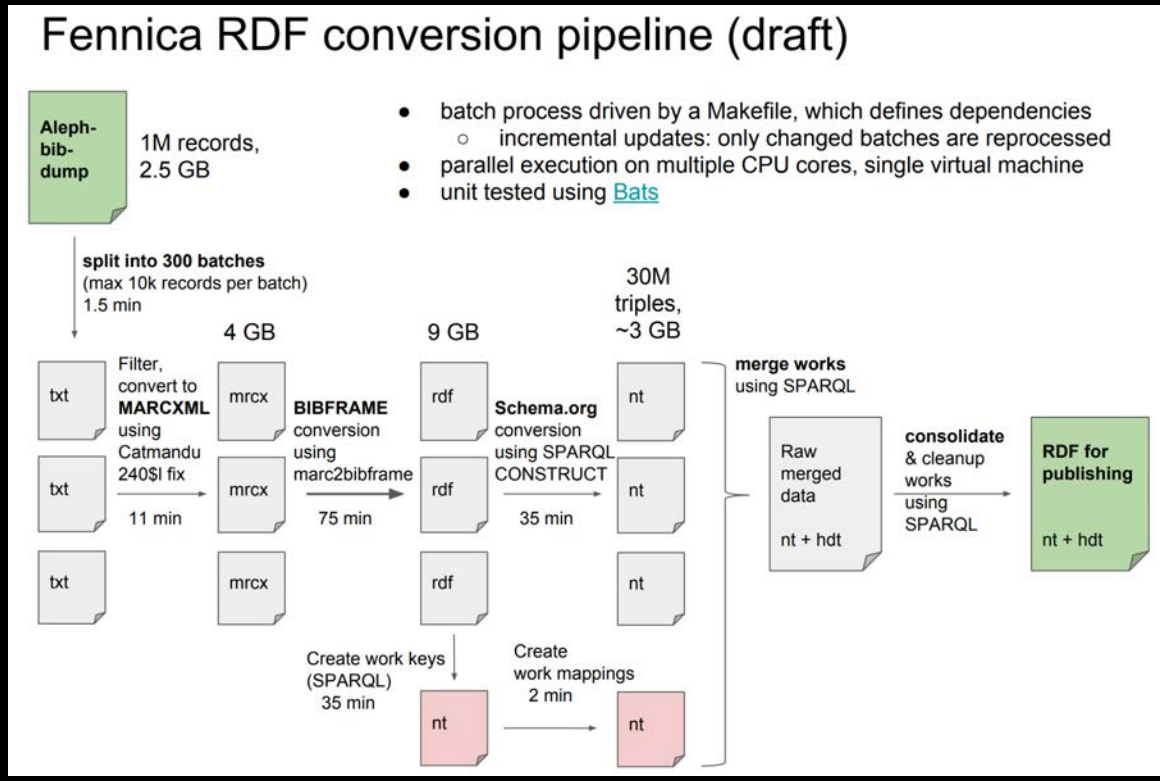
- add URIs first
- establish 2-way conversions for import/export



https://goo.gl/vwUiJY

National Library of Finland

- MARC to BIBFRAME to schema.org
- Focus on web publication, hence schema.org



http://swib.org/swib16/slides/suominen_silos.pdf

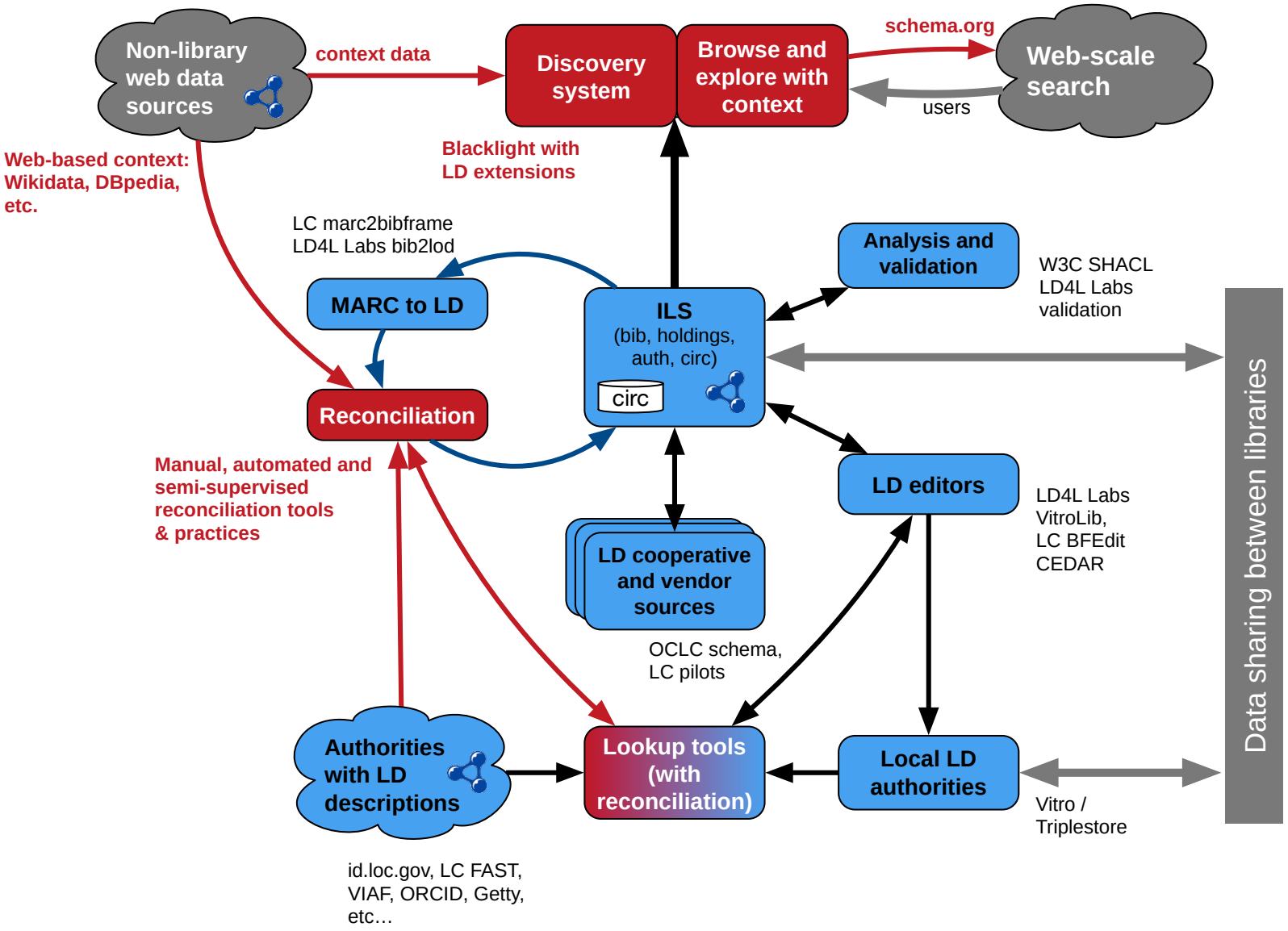
How close are we to linked data catalogs?

Let's not forget utility

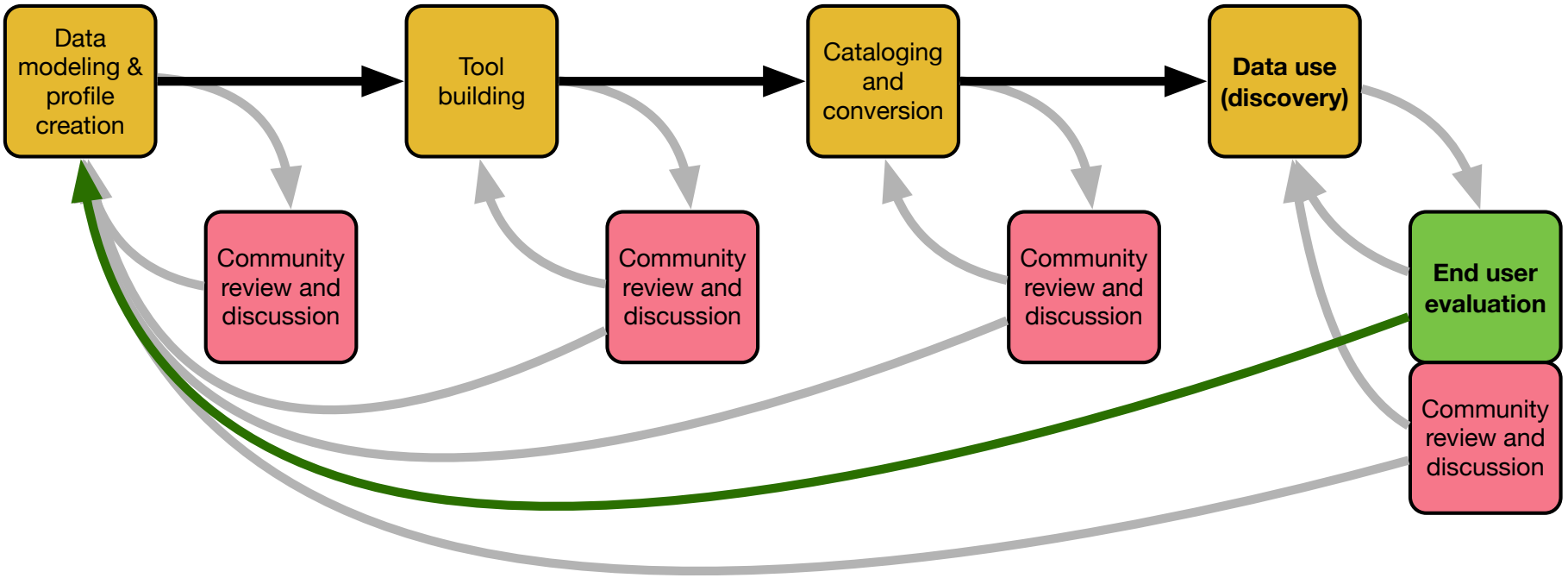
"Catalogers are primarily concerned about the quality and consistency of the data they produce, while technologists are primarily concerned with the techniques and tools that can be used to manipulate it."

[Jeff Edmunds, https://scholarsphere.psu.edu/concern/generic_works/44558d45t]

Linked Data catalog ecosystem



Catalog system feedback cycles



Open Collaborations (around software)

Free and Open Source Software

“Over The Wall”

- Simply make a copy of the source code available
- Exemplified by many uses of SourceForge (though has more features)
- Sharing but not collaboration

... better than *not* sharing

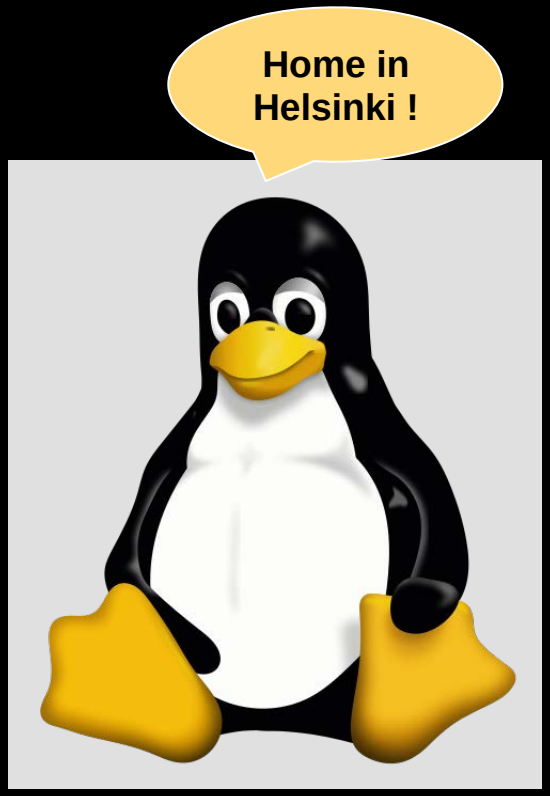
Open Development

- and related: “Social Coding”
- Share changes as they are made and provide means of contact/input
- Exemplified by basic use of GitHub (other services too)
- License for re-use

better than
“Over The Wall”

Community Development

- aka “Community Source Software”
- Multiple parties working together toward shared goals
- Norms
- Coordination
- Governance



https://commons.wikimedia.org/wiki/File:Tux.svg

https://commons.wikimedia.org/wiki/File:Apache_Software_Foundation_Logo_(2016).svg

Apache 2.0 License

Samvera (formerly Hydra)

- Framework and “solution bundles” for repository and DAM systems
- Blacklight/Solr + Fedora + Ruby
- 30+ partner institutions

- Vibrant and supportive community
- Yearly conference and other meetings
- Training
- Currently considering stronger governance options

https://samvera.org/

International Image Interoperability Framework

"A community of the world's leading libraries and image repositories working to produce a community framework and interoperable technology for image delivery."

- Primary outputs are specifications, software developed by sub-groups
- IIIF Consortium formed in 2015 to support growth and adoption
  - > 40 members, growing rapidly
  - Memberships pay for staff (2)
  - Libraries, museums, galleries, vendors

http://iiif.io/

Final thoughts

Most of interesting big challenges require collaboration to realize, including the ones I've mentioned:

- opening access to scholarly literature, making it discoverable, and linking researchers to their contributions
- moving to the next generation of library catalogs better integrated with the web

Kiitos!

@zimeon
simeon.warner@cornell.edu