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
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)

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...
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)
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
Clicks from SPIRES to DOI links and arXiv in October 2008

Clicks to arXiv: 82%

Clicks to journal websites: 18%
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

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

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
• ...

Cori Bargmann
Like This Page • April 26

I am thrilled to announce that the Chan Zuckerberg Initiative will be collaborating with Cold Spring Harbor Laboratory to help support their online platform, bioRxiv. This free service enables life science researchers to share drafts of their papers — known as preprints — before they appear in journals.

Expanded access to these drafts can dramatically accelerate the pace of discovery, and in turn, our understanding of health and disease.

In biology, it can take months, and sometimes years, for a manuscript to be peer-reviewed and published. That means scientists often don’t see important findings until long after the original discoveries. Preprint repositories like bioRxiv complement, and even augment, the peer-review process by making these results available instantly so they can be discussed, shared, and improved upon by the entire scientific community before publication.
New abcXiv and acquisitions

OSF Preprints

Providers

- OSF Preprints (1,432)
- AgrXiv (17)
- arXiv (1,254,157)
- bioRxiv (15,400)
- BITSS (13)
- Cogprints (284)
- EarthArXiv (9)
- engrXiv (125)
- FocUS Archive (2)
- INA-Rxiv (672)
- LawArXiv (365)
- LIS Scholarship Archive (25)
- MindRxiv (22)
- NutriXiv (11)
- PaleorXiv (39)
- PeerJ (2,580)
- Preprints.org (2,454)
- PsyArXiv (970)
- RePEc (803,492)
- SocArXiv (1,565)
- SportRxiv (11)
- Thesis Commons (82)

Elsevier buys research sharing website

SSRN acquisition angers some academics who say group may restrict access
Overlap & competition

Announcement of new Economics (econ) archive

September 26, 2017

Dear Colleagues,

An Economics section of the scientific repository arXiv is opening this month. arXiv is internationally acknowledged as a pioneering open access preprint repository. It has transformed the scholarly communication infrastructure of multiple fields of physics and plays an increasingly prominent role in mathematics, computer science, quantitative biology, quantitative finance, and statistics. arXiv is an essential component of scientific communication for many researchers worldwide in order to rapidly and widely disseminate their findings, establish priority of their discoveries, and seek feedback to help improve their work. It is hosted by the Cornell University Library with additional funding from 220 members libraries and several scientific foundations including the Simons Foundation.

The Economics arXiv will start with a single subject area of Econometrics, but according to demand we expect to add more subject areas in the future, and the Econometrics subject can also be subdivided further. The reason to start with Econometrics is that a significant number of Econometricians already regularly submit their work to the statistics domain of the arXiv. We hope that creating an explicit Econometrics domain will result in a majority of Econometricians to submit their work to the arXiv, so that most of the new papers in the field are disseminated in that way. If that goal is achieved, then regularly checking the Econometrics arXiv will become an efficient and easy way to stay current on the Econometrics literature. We initially want to focus on methodological papers, in agreement with the following description:

- econ.EM (econometrics): Econometric Theory, Micro-Econometrics, Macro-Econometrics, Empirical Content of Economic Relations discovered via New Methods, Methodological Aspects of the Application of Statistical Inference to Economic Data
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

Resource - Item - Record

resource

item

item has identifier

all available metadata about this sculpture

records

Dublin Core metadata

MARC21 metadata

branding metadata

record has identifier + metadata format + datestamp
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
ResourceSync: Leveraging Sitemaps for Resource Synchronization

Bernhard Haslhofer, Simeon Warner, Carl Lagoze, Martin Klein, Robert Sanderson, Michel 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.
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."
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

Sitemaps +
• multiple sets
• fixity
• links
• changes only
• dumps

Also supports Notifications (push) as optional extension
Aggregating the world’s open access research papers

We offer seamless access to millions of open access research papers, enrich the collected data for text-mining and provide unique services to the research community.
Use Case 3: Replace OAI-PMH with ResourceSync

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

Key publishers (OA + hybrid OA)
- Elsevier
- Wiley-Blackwell
- Springer
- Taylor & Francis
- + many others

A range of bespoke APIs

Publisher connector

ResourceSync

CORE

ResourceSync

Mostly OAI-PMH

OA Repositories

OA Journals

Tested with resync client. 20 x 25MB sitemaps, 1M items ✔
The repository community should agree on a common new approach to harvesting. ResourceSync was designed to meet this need.
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](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
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

**VIAF**
- Scope: ?M
- Now: 6M

**ISNI**
- Scope: ?M
- Now: 9M

**ORCID**
- Scope: 8-20M active, +2-4M/year
- Now: 3.2M
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?
- + >> 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

- Links to other identities – leverage overlaps
- Biography and information shown under my control
- Sources indicated
- Researcher can choose what appears on their record
ORCID iD use

• 7000 journals use ORCID iDs, over 1500 of which require use by corresponding authors

• Researcher support from surveys:
  o 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
  o 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

More than 550 integrations across all sectors of the research community

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

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 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
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”

Linked Data

• A practical “semantic 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

Extensions

ArtFrame | Community adoption & revision?
RareMat | Community adoption & revision?
...others... | Community adoption & revision?
bfhc extension | ???

BIBFRAME1.0 | BIBFRAME2.0 | BIBFRAME3.x? | BIBFRAME4.x?
LD4L critique | bibliotek-o

Time

NOW
LD4L & 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


• 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
  o copy cataloging
  o 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
National Library of Finland

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

Fennica RDF conversion pipeline (draft)

- batch process driven by a Makefile, which defines dependencies
  - incremental updates: only changed batches are reprocessed
- parallel execution on multiple CPU cores, single virtual machine
- unit tested using Bats

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

Discovery system

Browse and explore with context

Web-scale search

Non-library web data sources

context data

Browse and explore with context

Blacklight with LD extensions

MARC to LD

ILS (bib, holdings, auth, circ)

LD cooperative and vendor sources

Analysis and validation

W3C SHACL
LD4L Labs validation

LD4L Labs bib2lod

LC marc2bibframe

reconciliation tools & practices

Manual, automated and semi-supervised reconciliation tools & practices

Web-based context: Wikidata, DBpedia, etc.

Data sharing between libraries

Authorities with LD descriptions

Lookup tools (with reconciliation)

Local LD authorities

Web-scale

Vitro / Triplestore

LD editors

LD4L Labs VitroLib, LC BFEdit CEDAR

LD cooperative and vendor sources

OCLC schema, LC pilots

Web search

Users

Authorities with LD descriptions

id.loc.gov, LC FAST, VIAF, ORCID, Getty, etc...

Linked Data catalog ecosystem
Catalog system feedback cycles

Data modeling & profile creation

Tool building

Cataloging and conversion

Data use (discovery)

Community review and discussion

Community review and discussion

Community review and discussion

End user evaluation

Community review and discussion
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

Home in Helsinki!

https://commons.wikimedia.org/wiki/File:Tux.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
  o > 40 members, growing rapidly
  o Memberships pay for staff (2)
  o 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