ResourceSync for Semantic Web Data Copying and Synchronization

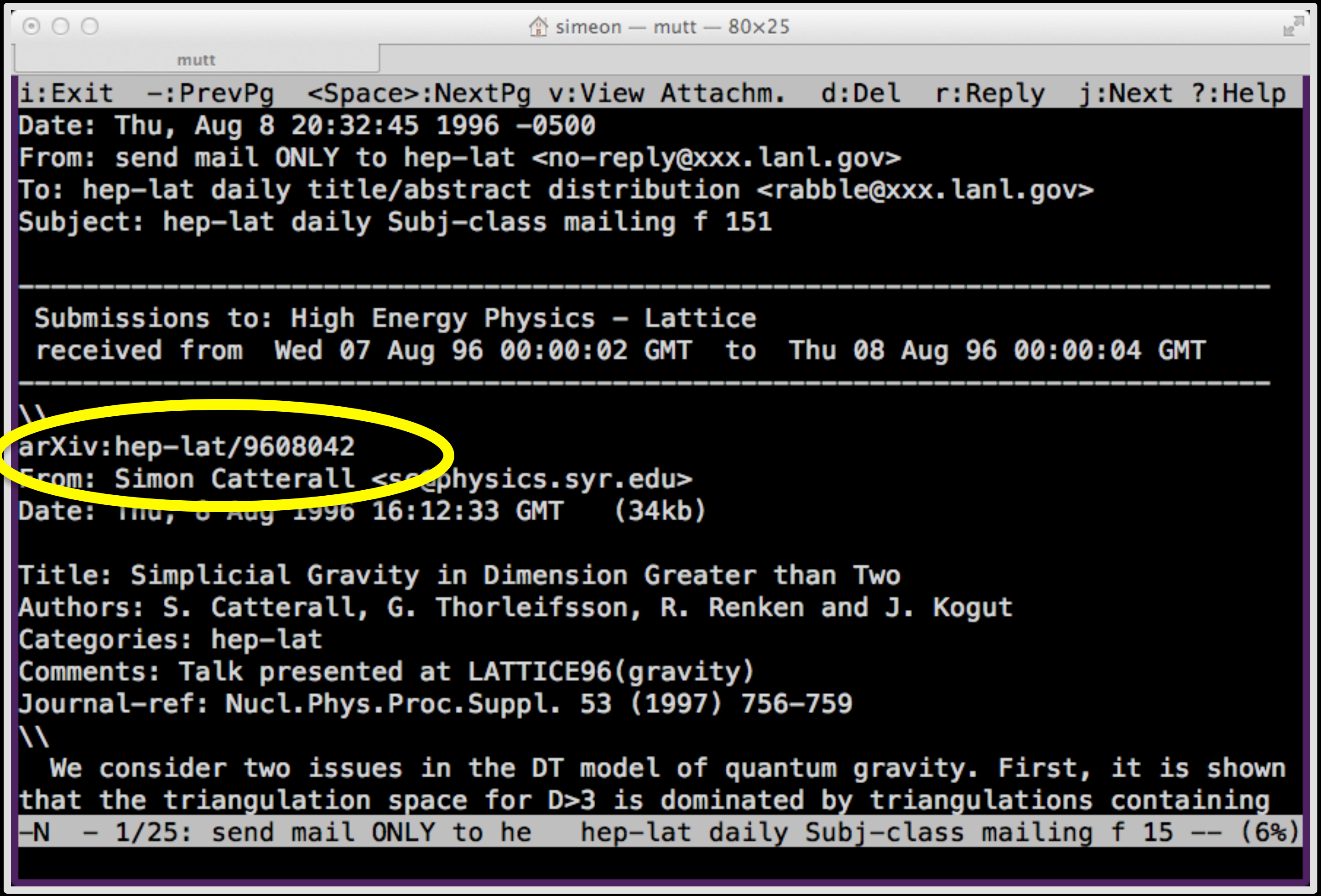
Simeon Warner (Cornell University)
http://orcid.org/0000-0002-7970-7855

SWIB13, Hamburg, Germany
2013-11-27

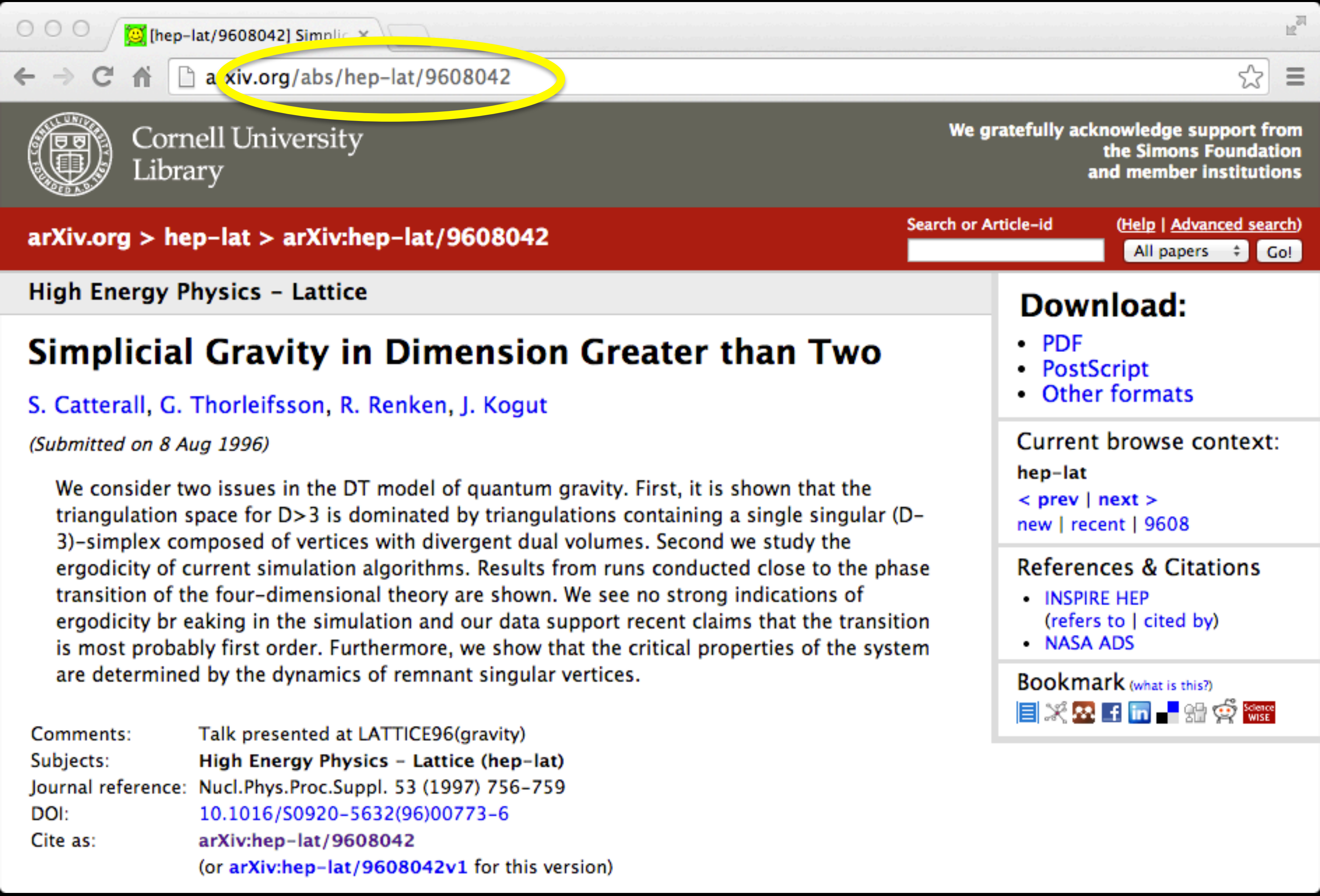
Menu

1. A personal spin
2. ResourceSync
    a. ResourceSync: Problem Perspective & Conceptual Approach
    b. Motivation & Use Cases
    c. Framework Walkthrough
    d. Framework Technical Details
    e. Implementation
3. ResourceSync and the Semantic Web

Typical morning, summer 1996



```
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%)
```



[hep-lat/9608042] Simplic
Cornell University Library
We gratefully acknowledge support from the Simons Foundation and member institutions
Search or Article-id
(Help | Advanced search)
All papers
Go!

arxiv.org/abs/hep-lat/9608042

arXiv.org > hep-lat > arXiv:hep-lat/9608042

High Energy Physics – Lattice

Simplicial Gravity in Dimension Greater than Two

S. Catterall, G. Thorleifsson, R. Renken, J. Kogut

(Submitted on 8 Aug 1996)

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 a single singular (D-3)-simplex composed of vertices with divergent dual volumes. Second we study the ergodicity of current simulation algorithms. Results from runs conducted close to the phase transition of the four-dimensional theory are shown. We see no strong indications of ergodicity br eaking in the simulation and our data support recent claims that the transition is most probably first order. Furthermore, we show that the critical properties of the system are determined by the dynamics of remnant singular vertices.

Comments:	Talk presented at LATTICE96(gravity)
Subjects:	**High Energy Physics – Lattice (hep-lat)**
Journal reference:	Nucl.Phys.Proc.Suppl. 53 (1997) 756-759
DOI:	10.1016/S0920-5632(96)00773-6
Cite as:	**arXiv:hep-lat/9608042** (or **arXiv:hep-lat/9608042v1** for this version)

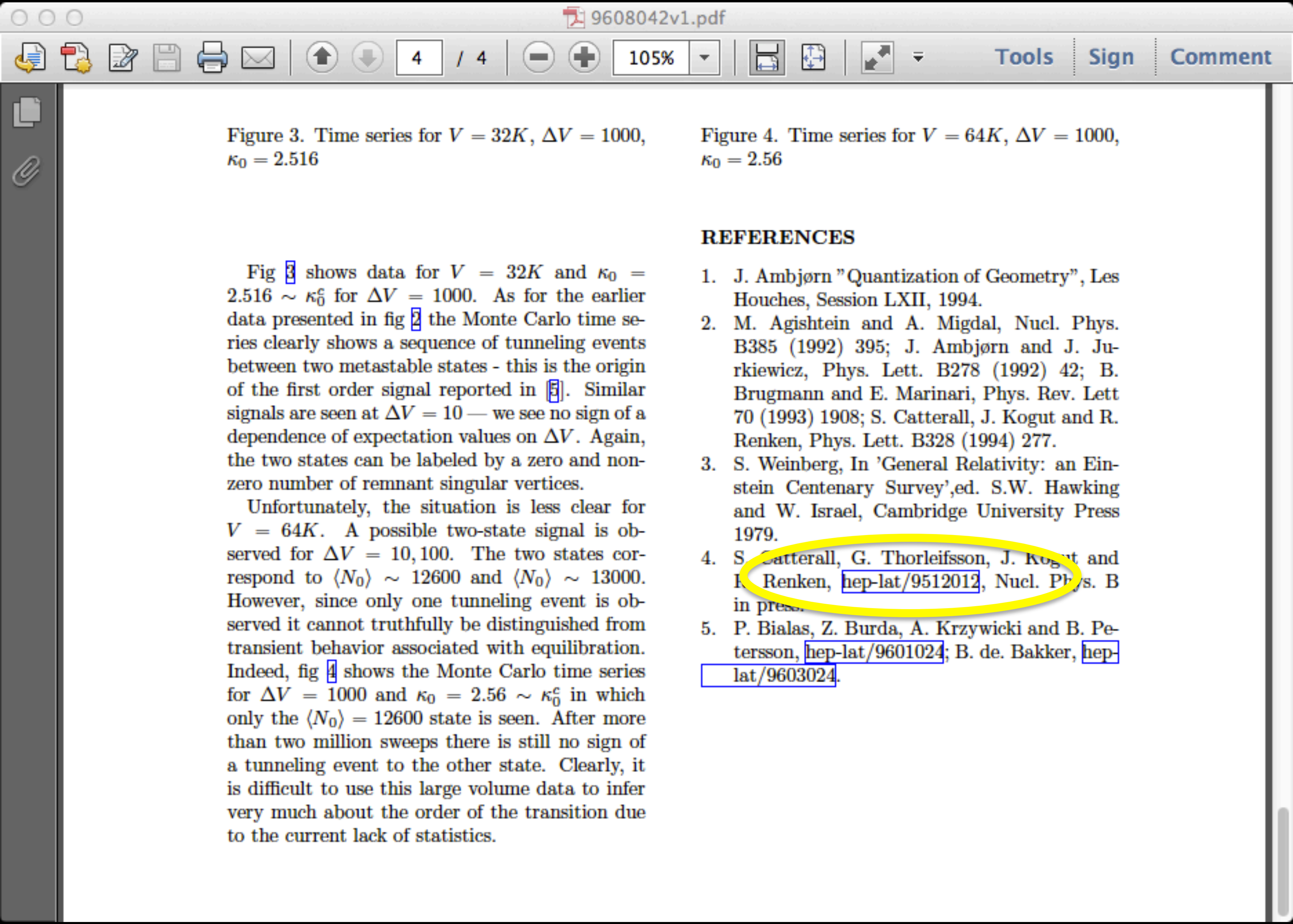


Figure 3. Time series for $V = 32K$, $\Delta V = 1000$, $\kappa_0 = 2.516$

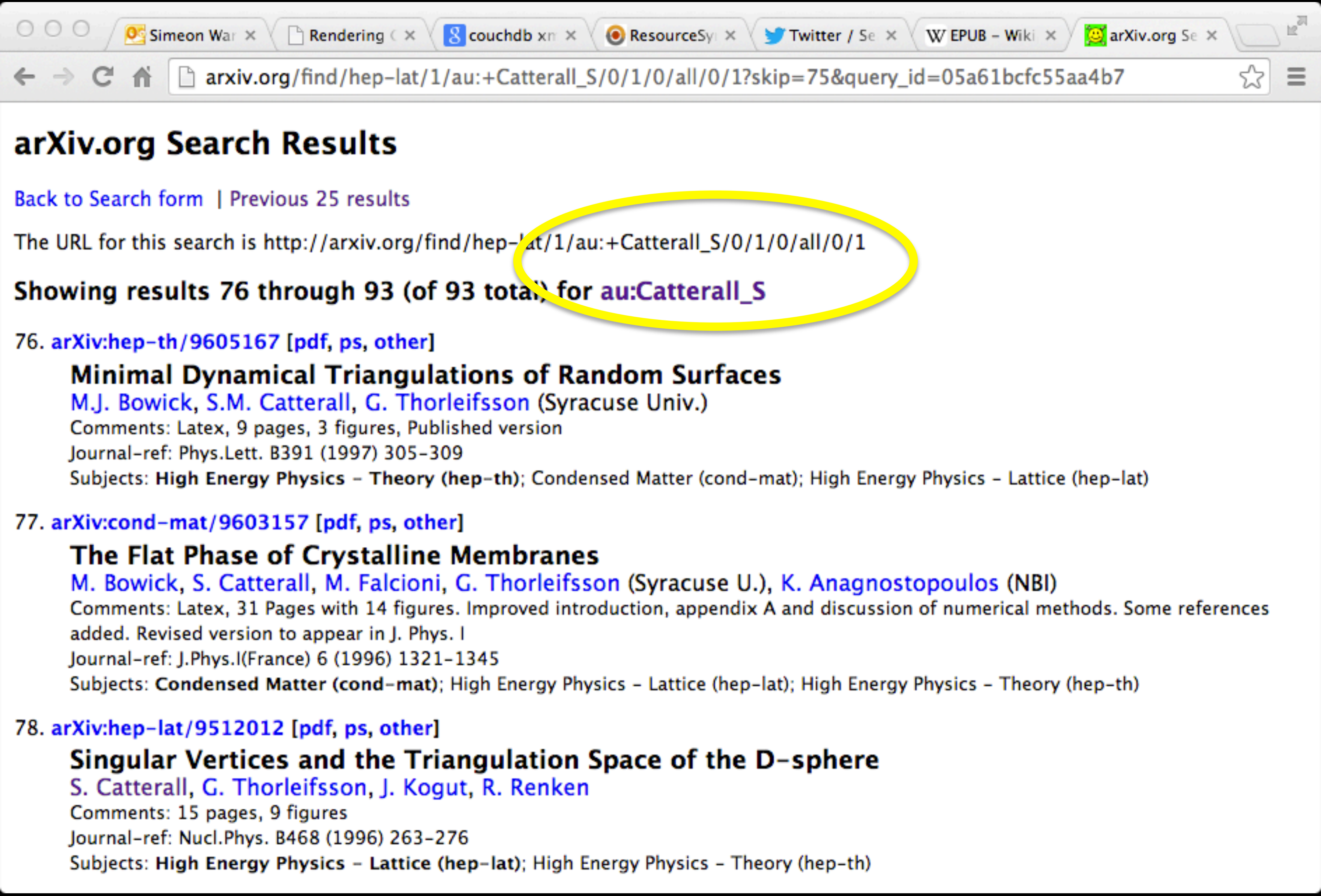
Figure 4. Time series for $V = 64K$, $\Delta V = 1000$, $\kappa_0 = 2.56$

Fig 3 shows data for $V = 32K$ and $\kappa_0 = 2.516 \sim \kappa_0^c$ for $\Delta V = 1000$. As for the earlier data presented in fig 2 the Monte Carlo time series clearly shows a sequence of tunneling events between two metastable states - this is the origin of the first order signal reported in [5]. Similar signals are seen at $\Delta V = 10$ — we see no sign of a dependence of expectation values on ΔV. Again, the two states can be labeled by a zero and non-zero number of remnant singular vertices.

Unfortunately, the situation is less clear for $V = 64K$. A possible two-state signal is observed for $\Delta V = 10, 100$. The two states correspond to $\langle N_0 \rangle \sim 12600$ and $\langle N_0 \rangle \sim 13000$. However, since only one tunneling event is observed it cannot truthfully be distinguished from transient behavior associated with equilibration. Indeed, fig 4 shows the Monte Carlo time series for $\Delta V = 1000$ and $\kappa_0 = 2.56 \sim \kappa_0^c$ in which only the $\langle N_0 \rangle = 12600$ state is seen. After more than two million sweeps there is still no sign of a tunneling event to the other state. Clearly, it is difficult to use this large volume data to infer very much about the order of the transition due to the current lack of statistics.

REFERENCES

1. J. Ambjørn "Quantization of Geometry", Les Houches, Session LXII, 1994.
2. M. Agishtein and A. Migdal, Nucl. Phys. B385 (1992) 395; J. Ambjørn and J. Jurkiewicz, Phys. Lett. B278 (1992) 42; B. Brugmann and E. Marinari, Phys. Rev. Lett 70 (1993) 1908; S. Catterall, J. Kogut and R. Renken, Phys. Lett. B328 (1994) 277.
3. S. Weinberg, In 'General Relativity: an Einstein Centenary Survey',ed. S.W. Hawking and W. Israel, Cambridge University Press 1979.
4. S. Catterall, G. Thorleifsson, J. Kogut and R. Renken, hep-lat/9512012, Nucl. Phys. B in press.
5. P. Bialas, Z. Burda, A. Krzywicki and B. Petersson, hep-lat/9601024; B. de. Bakker, hep-lat/9603024.


Simeon War
Rendering (
couchdb xn
ResourceSy
Twitter / Se
EPUB – Wiki
arXiv.org Se
arxiv.org/find/hep-lat/1/au:+Catterall_S/0/1/0/all/0/1?skip=75&query_id=05a61bcfc55aa4b7

arXiv.org Search Results

Back to Search form | Previous 25 results

The URL for this search is http://arxiv.org/find/hep-lat/1/au:+Catterall_S/0/1/0/all/0/1

Showing results 76 through 93 (of 93 total) for au:Catterall_S

76. arXiv:hep-th/9605167 [pdf, ps, other]

Minimal Dynamical Triangulations of Random Surfaces
M.J. Bowick, S.M. Catterall, G. Thorleifsson (Syracuse Univ.)
Comments: Latex, 9 pages, 3 figures, Published version
Journal-ref: Phys.Lett. B391 (1997) 305-309
Subjects: **High Energy Physics - Theory (hep-th)**; Condensed Matter (cond-mat); High Energy Physics - Lattice (hep-lat)

77. arXiv:cond-mat/9603157 [pdf, ps, other]

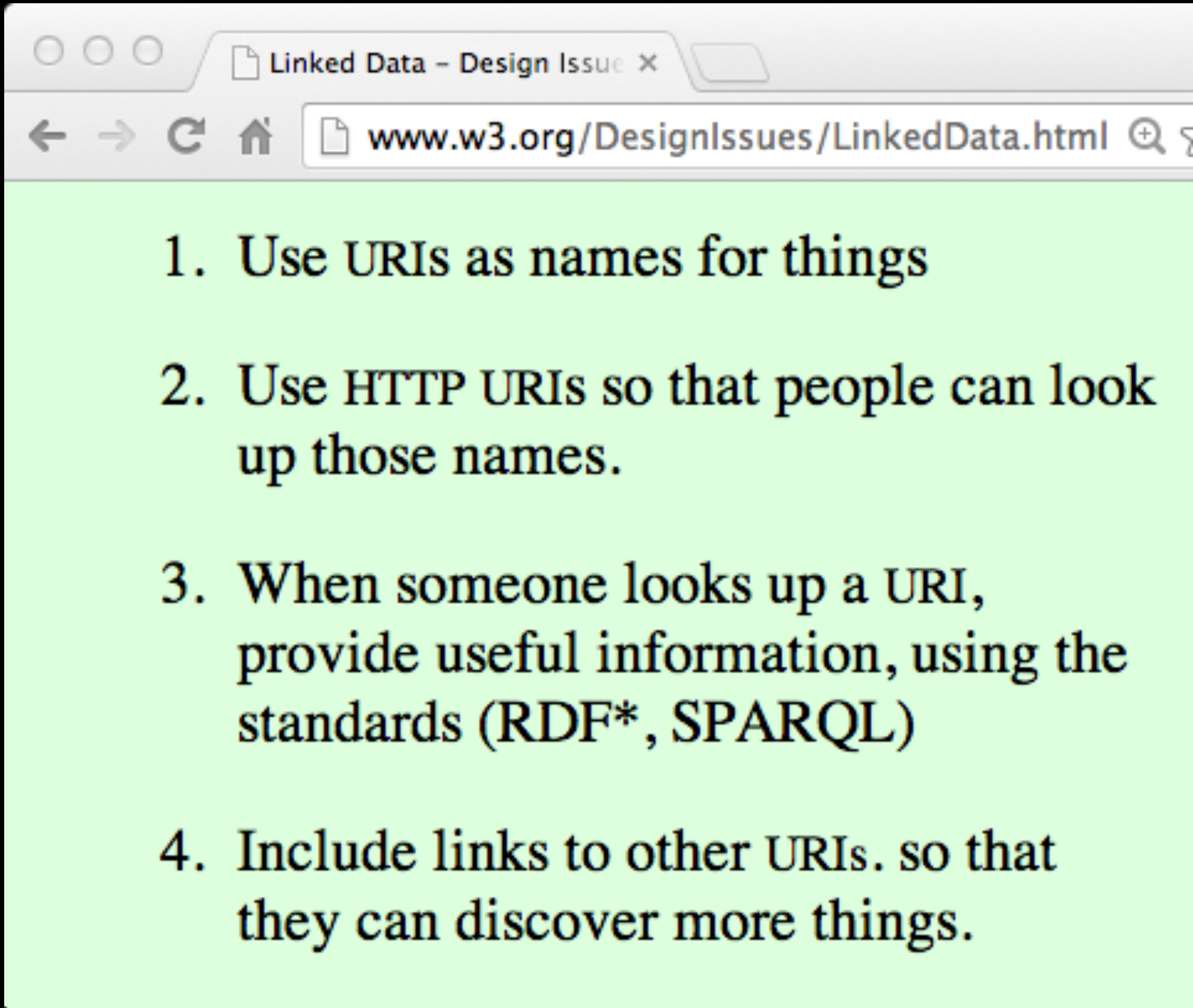
The Flat Phase of Crystalline Membranes
M. Bowick, S. Catterall, M. Falcioni, G. Thorleifsson (Syracuse U.), K. Anagnostopoulos (NBI)
Comments: Latex, 31 Pages with 14 figures. Improved introduction, appendix A and discussion of numerical methods. Some references added. Revised version to appear in J. Phys. I
Journal-ref: J.Phys.I(France) 6 (1996) 1321-1345
Subjects: **Condensed Matter (cond-mat)**; High Energy Physics - Lattice (hep-lat); High Energy Physics - Theory (hep-th)

78. arXiv:hep-lat/9512012 [pdf, ps, other]

Singular Vertices and the Triangulation Space of the D-sphere
S. Catterall, G. Thorleifsson, J. Kogut, R. Renken
Comments: 15 pages, 9 figures
Journal-ref: Nucl.Phys. B468 (1996) 263-276
Subjects: **High Energy Physics - Lattice (hep-lat)**; High Energy Physics - Theory (hep-th)

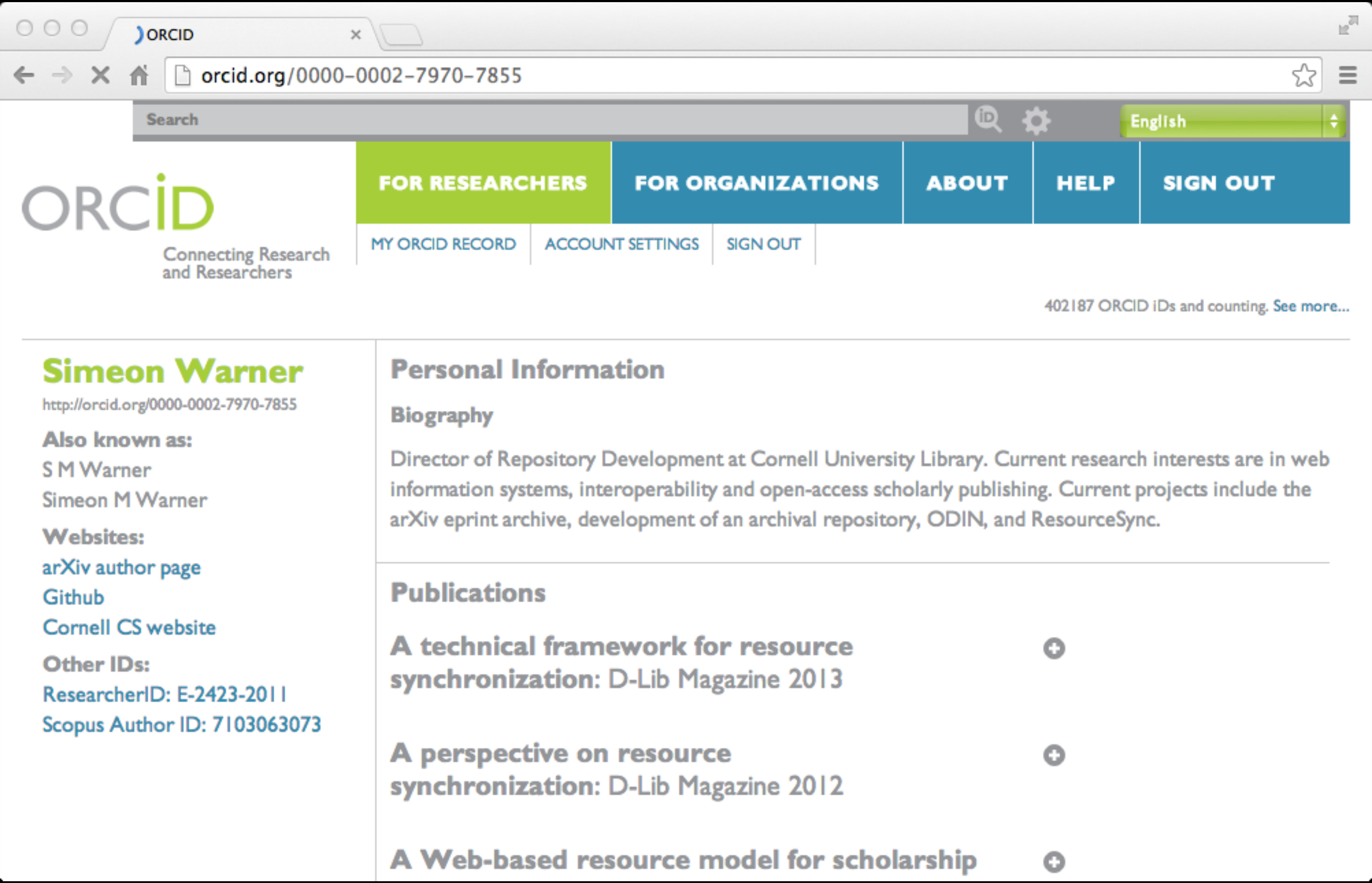
Linked world – but no data

1. Names for articles, people
2. HTTP to get data
3. *(no machine data)*
4. Have links to other things



Linked Data – Design Issue
www.w3.org/DesignIssues/LinkedData.html

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.



402187 ORCID iDs and counting. See more...

Simeon Warner

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

Also known as:

S M Warner

Simeon M Warner

Websites:

arXiv author page

Github

Cornell CS website

Other IDs:

ResearcherID: E-2423-2011

Scopus Author ID: 7103063073

Personal Information

Biography

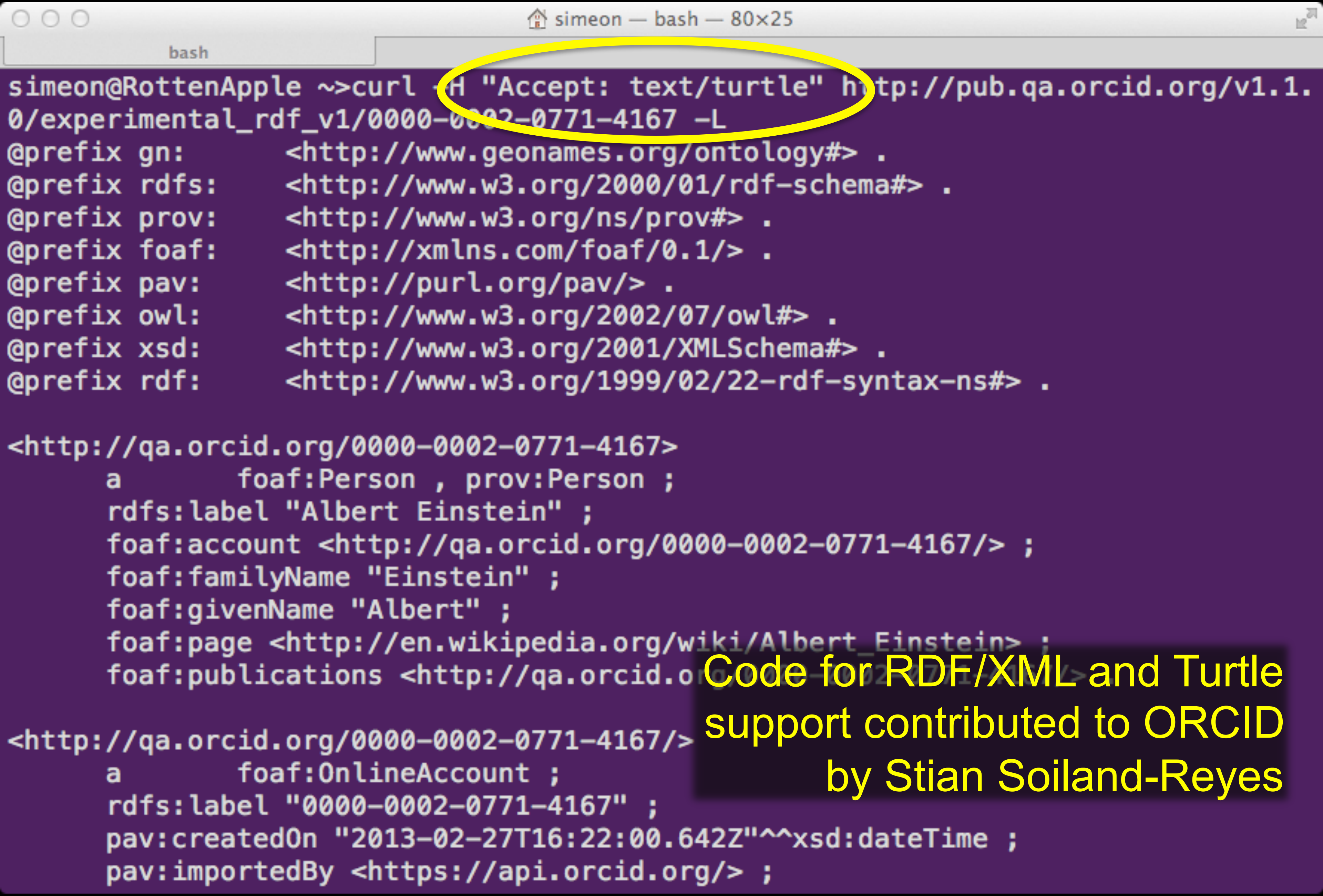
Director of Repository Development at Cornell University Library. Current research interests are in web information systems, interoperability and open-access scholarly publishing. Current projects include the arXiv eprint archive, development of an archival repository, ODIN, and ResourceSync.

Publications

A technical framework for resource synchronization: D-Lib Magazine 2013

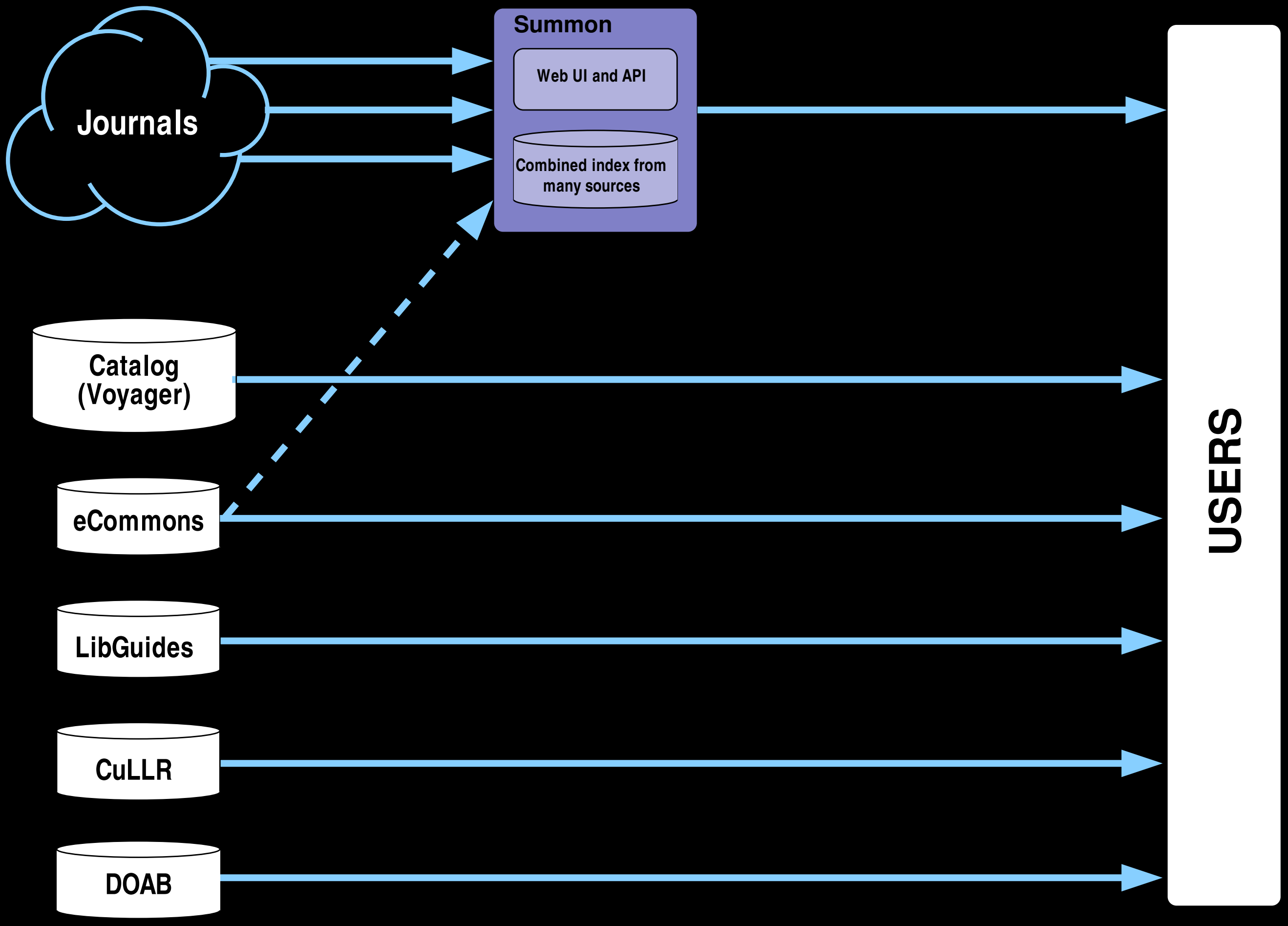
A perspective on resource synchronization: D-Lib Magazine 2012

A Web-based resource model for scholarship

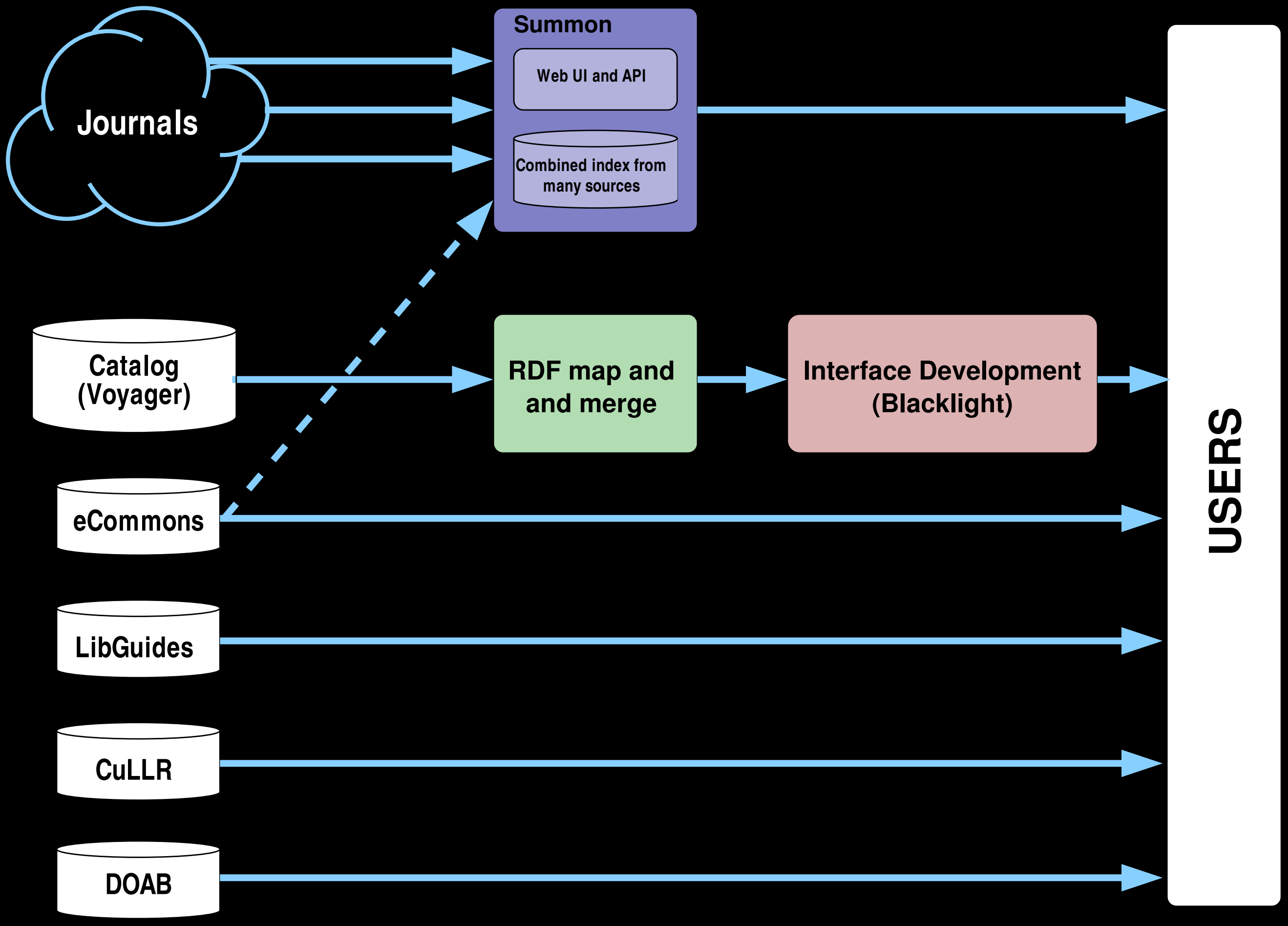


```
simeon@RottenApple ~>curl -H "Accept: text/turtle" http://pub.qa.orcid.org/v1.1.
0/experimental_rdf_v1/0000-0002-0771-4167 -L
@prefix gn:     <http://www.geonames.org/ontology#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix prov:   <http://www.w3.org/ns/prov#> .
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
@prefix pav:    <http://purl.org/pav/> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://qa.orcid.org/0000-0002-0771-4167>
      a       foaf:Person , prov:Person ;
      rdfs:label "Albert Einstein" ;
      foaf:account <http://qa.orcid.org/0000-0002-0771-4167/> ;
      foaf:familyName "Einstein" ;
      foaf:givenName "Albert" ;
      foaf:page <http://en.wikipedia.org/wiki/Albert_Einstein> ;
      foaf:publications <http://qa.orcid.o

<http://qa.orcid.org/0000-0002-0771-4167/>
      a       foaf:OnlineAccount ;
      rdfs:label "0000-0002-0771-4167" ;
      pav:createdOn "2013-02-27T16:22:00.642Z"^^xsd:dateTime ;
      pav:importedBy <https://api.orcid.org/> ;
```

Code for RDF/XML and Turtle support contributed to ORCID by Stian Soiland-Reyes



Discovery at Cornell



Journals

Summon

Web UI and API

Combined index from many sources

Catalog (Voyager)

RDF map and and merge

Interface Development (Blacklight)

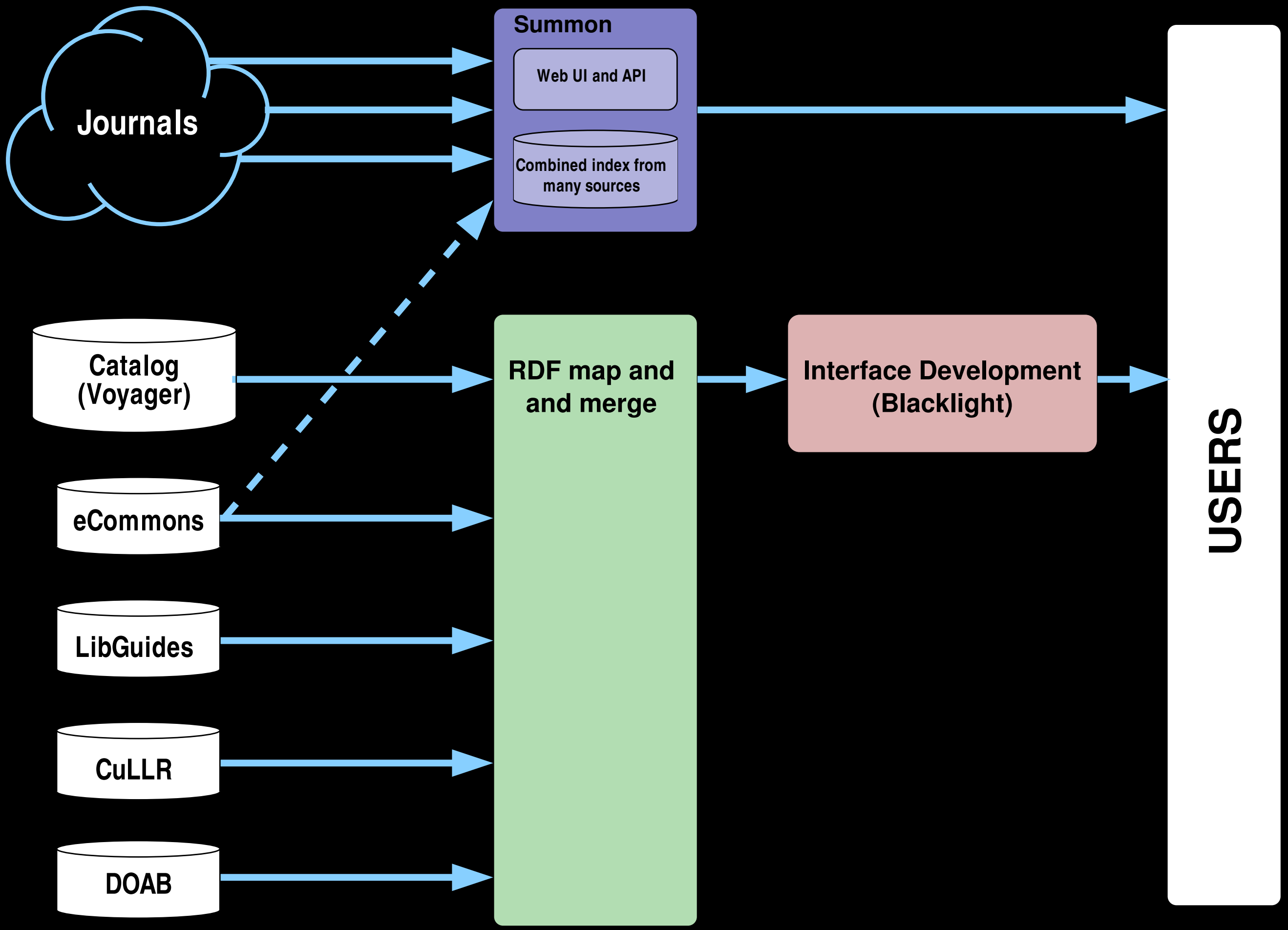
eCommons

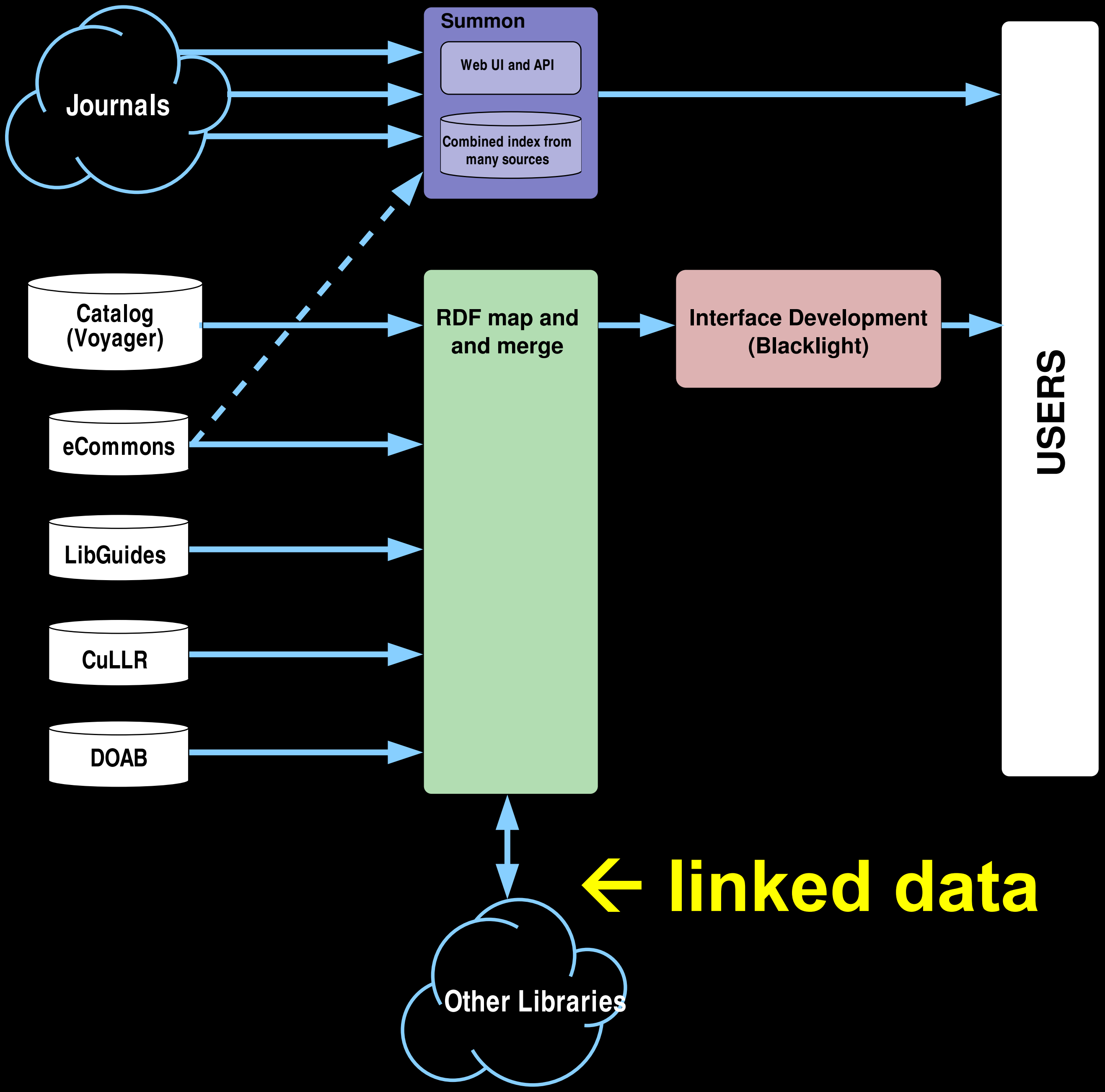
LibGuides

CuLLR

DOAB

USERS


Journals
Summon
Web UI and API
Combined index from many sources
Catalog (Voyager)
eCommons
LibGuides
CuLLR
DOAB
RDF map and and merge
Interface Development (Blacklight)
USERS



Journals

Summon

Web UI and API

Combined index from many sources

Catalog (Voyager)

eCommons

LibGuides

CuLLR

DOAB

RDF map and and merge

Interface Development (Blacklight)

USERS

Other Libraries

← linked data

ResourceSync:

A Web-Based Resource Synchronization Framework

These following slides are excerpted from the ResourceSync tutorial. The most recent version of the full tutorial slides is available at http://www.slideshare.net/OpenArchivesInitiative/resourcesync-tutorial

#resourcesync

ResourceSync is funded by
The Sloan Foundation & JISC

OAI

Herbert Van de Sompel
Martin Klein
Robert Sanderson
(Los Alamos National Laboratory)

Simeon Warner
(Cornell University)

Berhard Haslhofer
(University of Vienna)

Michael L. Nelson
(Old Dominion University)

Carl Lagoze
(University of Michigan)

NISO

Todd Carpenter
Nettie Lagace

University of Oxford

Graham Klyne

Lyrasis

Peter Murray

ResourceSync Technical Group

Ex Libris Inc.

Shlomo Sanders

JISC

Paul Walk

Richard Jones

Stuart Lewis

LOCKSS

David Rosenthal

RedHat

Christian Sadilek

Library of Congress

Kevin Ford

OCLC

Jeff Young

Timeline, Status of Specification(s)

- August 2013
  - Release of ResourceSync framework Core specification
    - Version 0.9.1
  - Public draft of ResourceSync Archives specification released
- September 2013
  - Core specification on its way to become an ANSI standard
- November 2013
  - Internal draft of ResourceSync Notification specification
- January 2014
  - Public draft of ResourceSync Notification specification
- Mid 2014
  - Core specification becomes ANSI/NISO standard

Menu

1. A personal spin
2. ResourceSync
    a. ResourceSync: Problem Perspective & Conceptual Approach
    b. Motivation & Use Cases
    c. Framework Walkthrough
    d. Framework Technical Details
    e. Implementation
3. ResourceSync and the Semantic Web

Synchronize What?

- Web resources
  - things with a URI that can be dereferenced
- Focus on needs of research communication and cultural heritage organizations but aim for generality

- Small websites/repositories (a few resources) to large repositories/datasets/linked data collections (many millions of resources)
- Low change frequency (weeks/months) to high change frequency (seconds)
- Synchronization latency and accuracy needs may vary

ResourceSync Problem

- Consider:
  - **Source** (server) A has resources that change over time: they get created, modified, deleted
  - **Destination** (servers) X, Y, and Z leverage (some) resources of Source A.
- Problem:
  - Destinations want to keep in step with the resource changes at Source A
- Goal:
  - Design an approach for resource synchronization aligned with the Web Architecture that has a fair chance of adoption by different communities.
    - The approach must scale better than recurrent HTTP HEAD/GET on resources.

Destination: Synchronization Needs

1. Baseline synchronization – A destination must be able to perform an initial load or catch-up with a source
    - avoid out-of-band setup

2. Incremental synchronization – A destination must have some way to keep up-to-date with changes at a source
    - subject to some latency; minimal: create/update/delete
    - allow to catch-up after destination has been offline

3. Audit – A destination should be able to determine whether it is synchronized with a source
    - regarding coverage and accuracy

Didn’t you sell us OAI-PMH?

Or... will ResourceSync replace OAI-PMH?

✓ Proven XML metadata transfer protocol
✓ Libraries in a number of programming languages
✓ Widely adopted *in our community*

x Predates REST, not “of the web”
x Not adopted for content transfer
x Technical issues with sets

- Devise a shared solution for data, metadata, linked data?

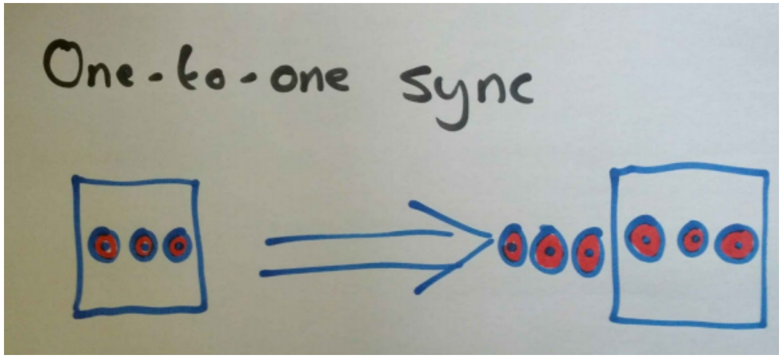
ResourceSync may replace, will likely coexistence

Menu

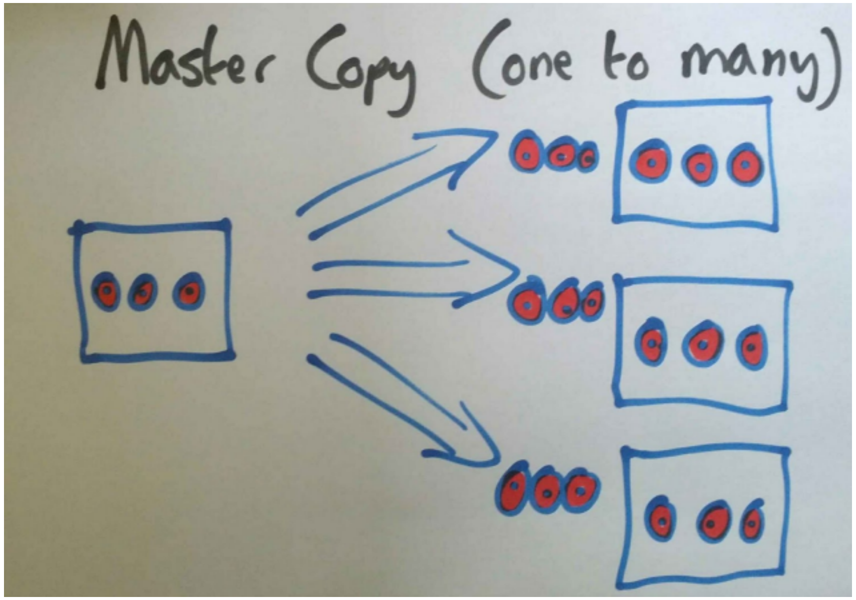
1. A personal spin
2. ResourceSync
   a. ResourceSync: Problem Perspective & Conceptual Approach
   b. Motivation & Use Cases
   c. Framework Walkthrough
   d. Framework Technical Details
   e. Implementation
3. ResourceSync and the Semantic Web

Use Cases – The Basics

a)

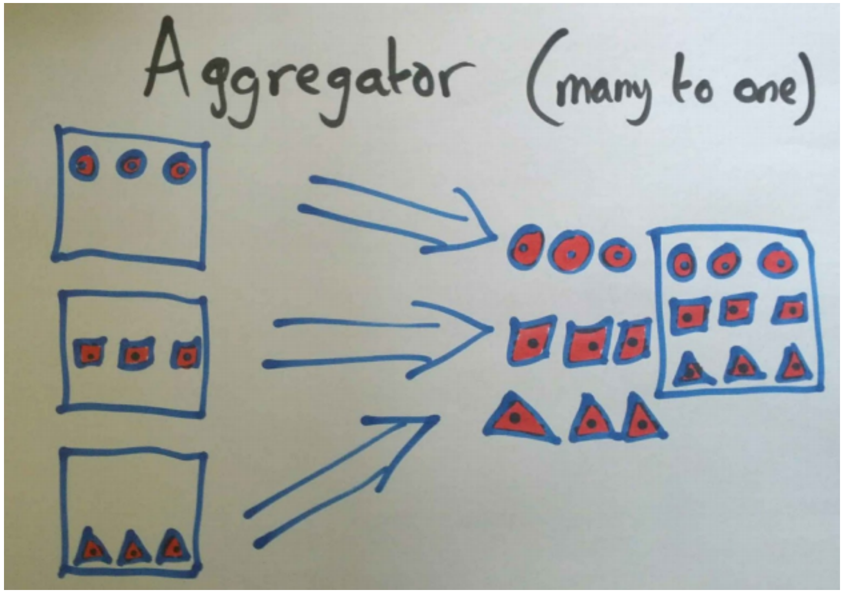


b)

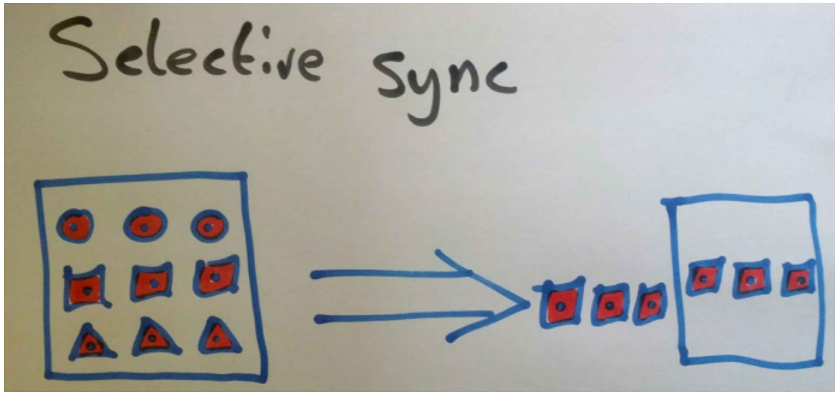


Use Cases – The Basics

c)

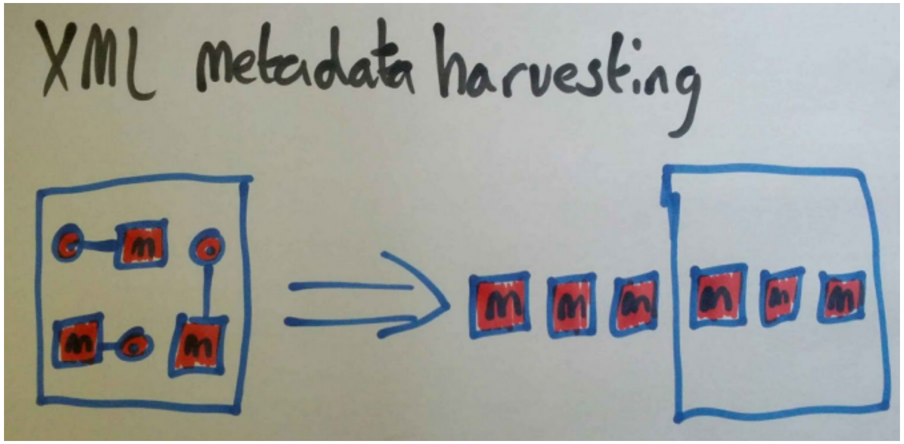


d)

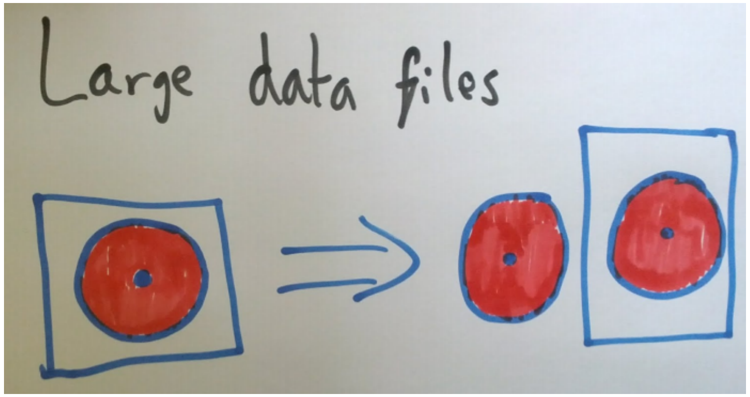


Use Cases – The not-so-Basics

e)



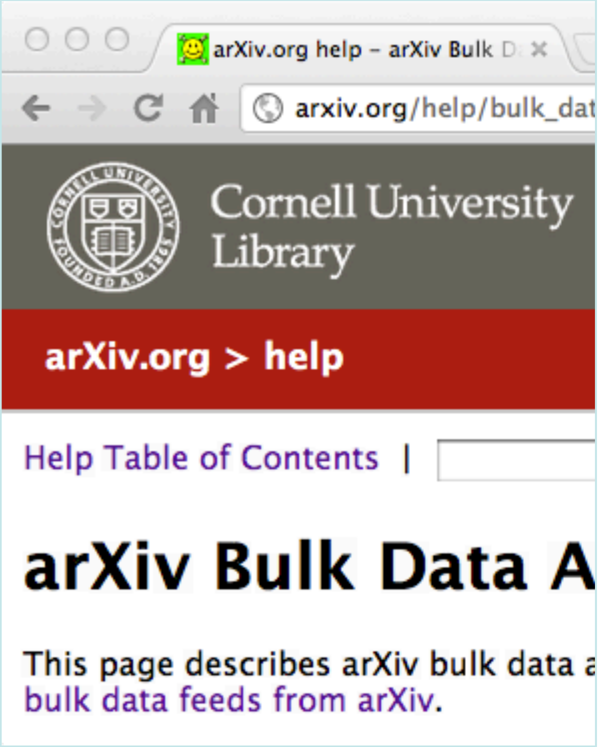
f)



Use Case 1: arXiv Mirroring and Data Sharing

- Repository of scholarly articles in physics, mathematics, computer science, etc.
- > 880k articles, ~1.5 revisions per article
- ~75k new articles per year
- metadata, source, PDF
- **~3.8M resources**
- **~2700 updates/day**

- Support
  - Mirroring
  - Sharing

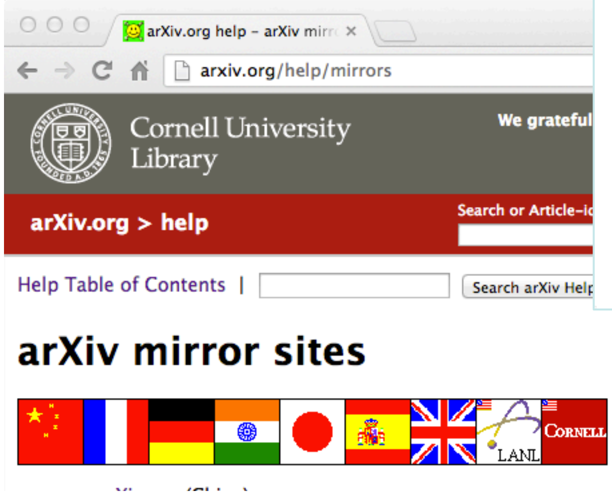


arXiv.org > help

Help Table of Contents |

arXiv Bulk Data A

This page describes arXiv bulk data a
bulk data feeds from arXiv.



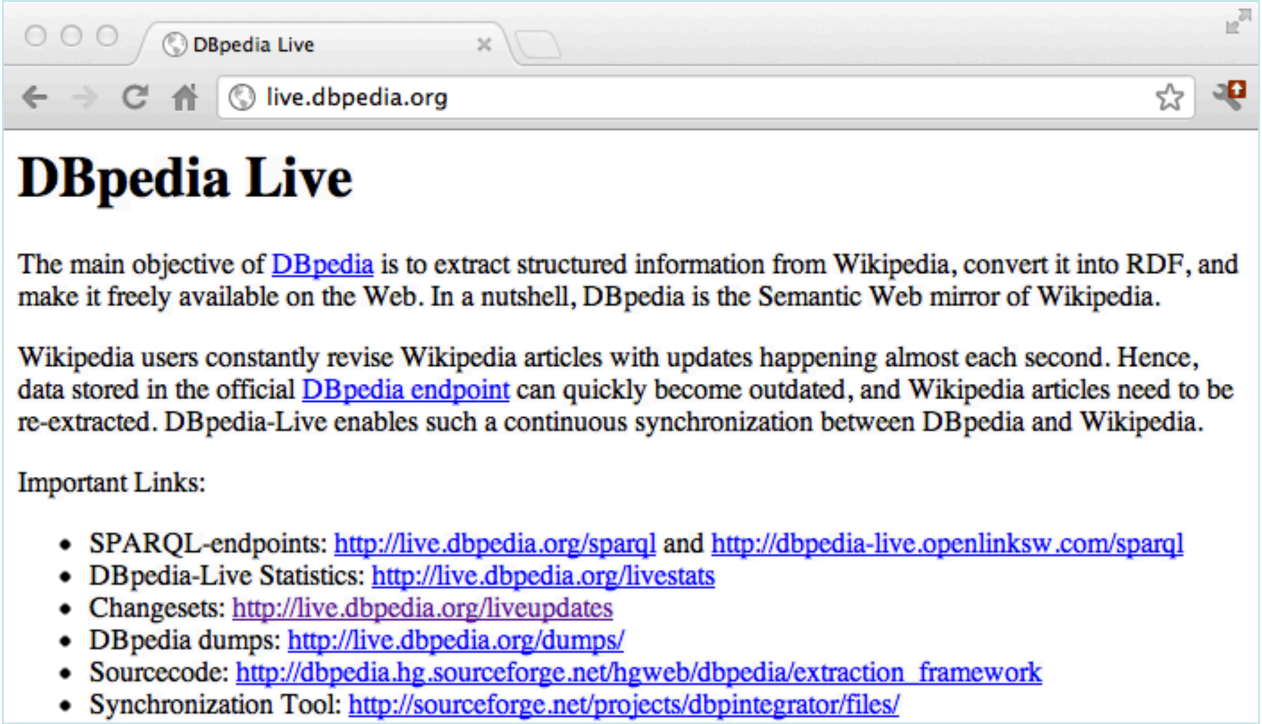
arXiv.org > help

Help Table of Contents |

arXiv mirror sites

Use Case 2: DBpedia Live Duplication

- Average of 2 updates per second
- Low latency desirable => need for a push technology



DBpedia Live

The main objective of DBpedia is to extract structured information from Wikipedia, convert it into RDF, and make it freely available on the Web. In a nutshell, DBpedia is the Semantic Web mirror of Wikipedia.

Wikipedia users constantly revise Wikipedia articles with updates happening almost each second. Hence, data stored in the official DBpedia endpoint can quickly become outdated, and Wikipedia articles need to be re-extracted. DBpedia-Live enables such a continuous synchronization between DBpedia and Wikipedia.

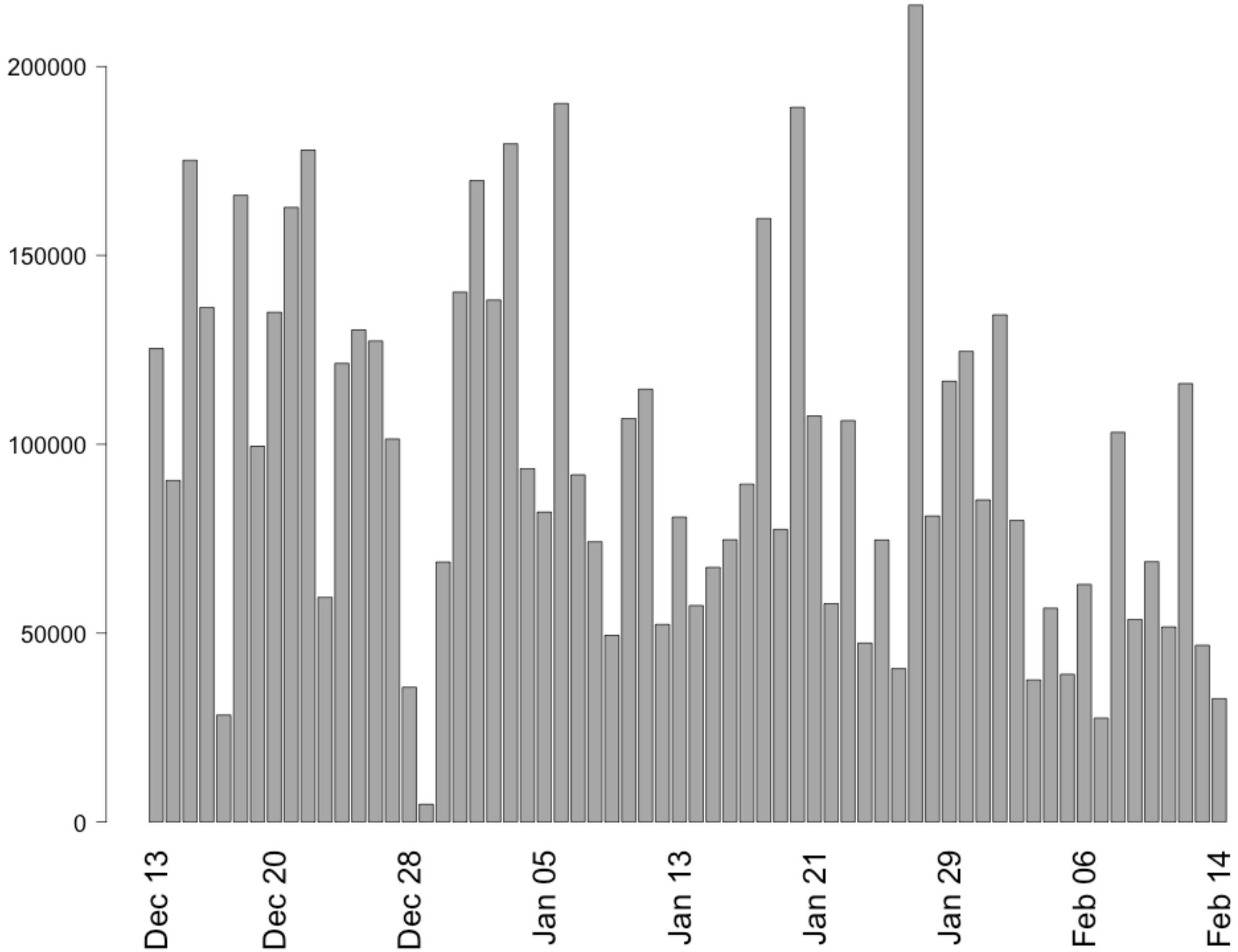
Important Links:

- SPARQL-endpoints: http://live.dbpedia.org/sparql and http://dbpedia-live.openlinksw.com/sparql
- DBpedia-Live Statistics: http://live.dbpedia.org/livestats
- Changesets: http://live.dbpedia.org/liveupdates
- DBpedia dumps: http://live.dbpedia.org/dumps/
- Sourcecode: http://dbpedia.hg.sourceforge.net/hgweb/dbpedia/extraction_framework
- Synchronization Tool: http://sourceforge.net/projects/dbpintegrator/files/

Use Case 2: DBpedia Live Duplication

- Daily traffic:
  - 99% updates
  - 0.6% deletions
  - 0.03% creations

- LANL experiments with push-based sync



Menu

1. A personal spin
2. ResourceSync
   1. ResourceSync: Problem Perspective & Conceptual Approach
   2. Motivation & Use Cases
   3. Framework Walkthrough
   4. Framework Technical Details
   5. Implementation
3. ResourceSync and the Semantic Web

Source: Core Synchronization Capabilities

P
U
L
L

1. Describing content – publish a list of resources available for synchronization to enable Destinations to perform an initial load or catch-up with a Source

2. Packaging content – bundle resources to enable bulk download by destinations

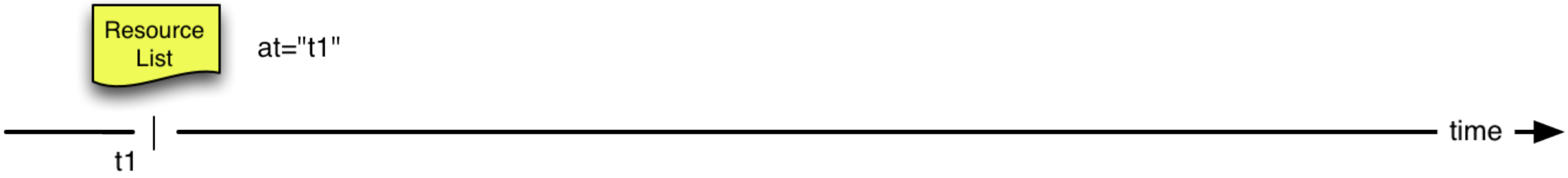
3. Describing changes – publish a list of resource changes to enable destinations to stay synchronized and decrease latency

4. Packaging changes – bundle resource changes for bulk download by destinations

Source Capability 1: Describing Content

In order to advertise the resources that a source wants destinations to know about, it may describe them:

- Publish a **Resource List**, a list of resource URIs and possibly associated metadata
  - Destination GETs the Resource List
  - Destination GETs listed resources by their URI
- A **Resource List** describes the state of a set of resources at one point in time (snapshot)

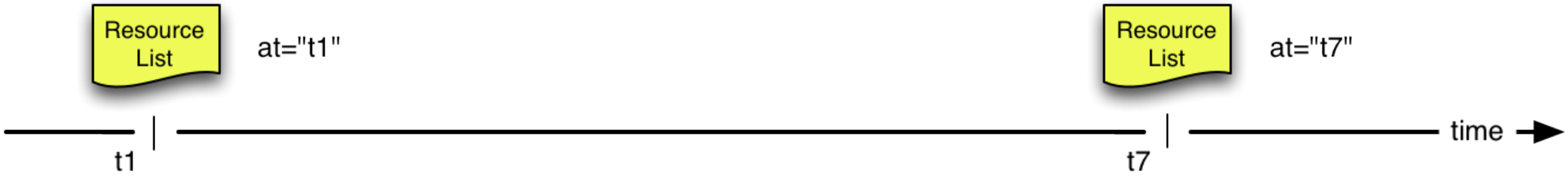


Resource List

at="t1"

t1

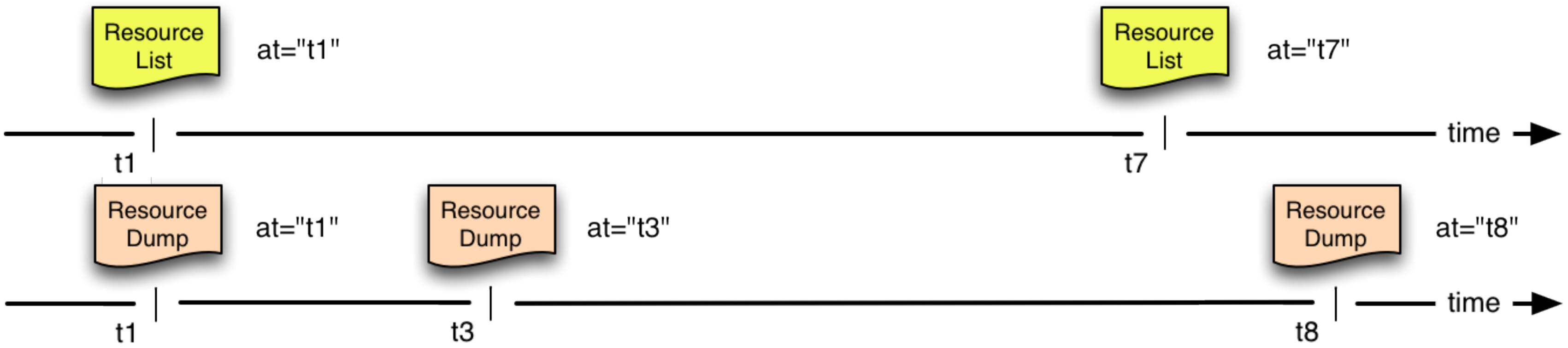
time


Resource List
at="t1"
Resource List
at="t7"
t1
t7
time

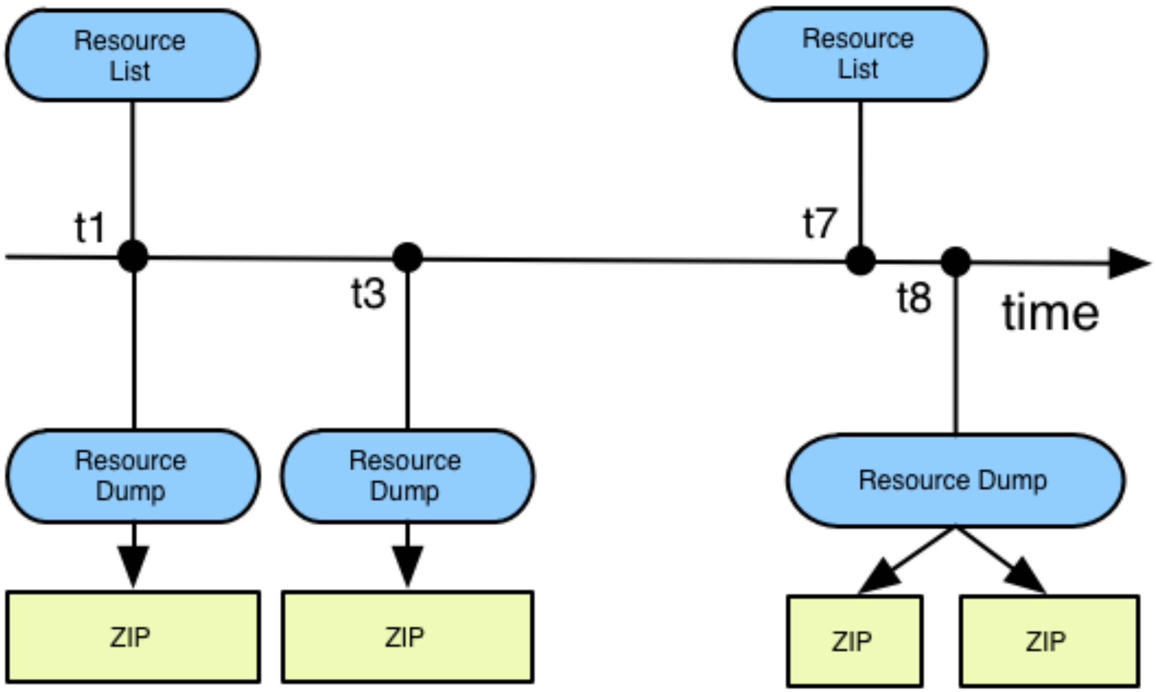
Source Capability 2: Packaging Content

By default, content is transferred in response to a GET issued by a destination against a URI of a source's resource. But a source may support additional mechanisms:

- Publish a **Resource Dump,** a document that points to packages of resource representations and necessary metadata
  - Destination GETs the package
  - Destination unpacks the package
  - ZIP format supported
- A **Resource Dump** and the packages it points to reflect the state of a set of resources at one point in time (snapshot)


Resource List
at="t1"
Resource List
at="t7"
t1
t7
time
Resource Dump
at="t1"
Resource Dump
at="t3"
Resource Dump
at="t8"
t1
t3
t8
time

Source:
Modular Capabilities



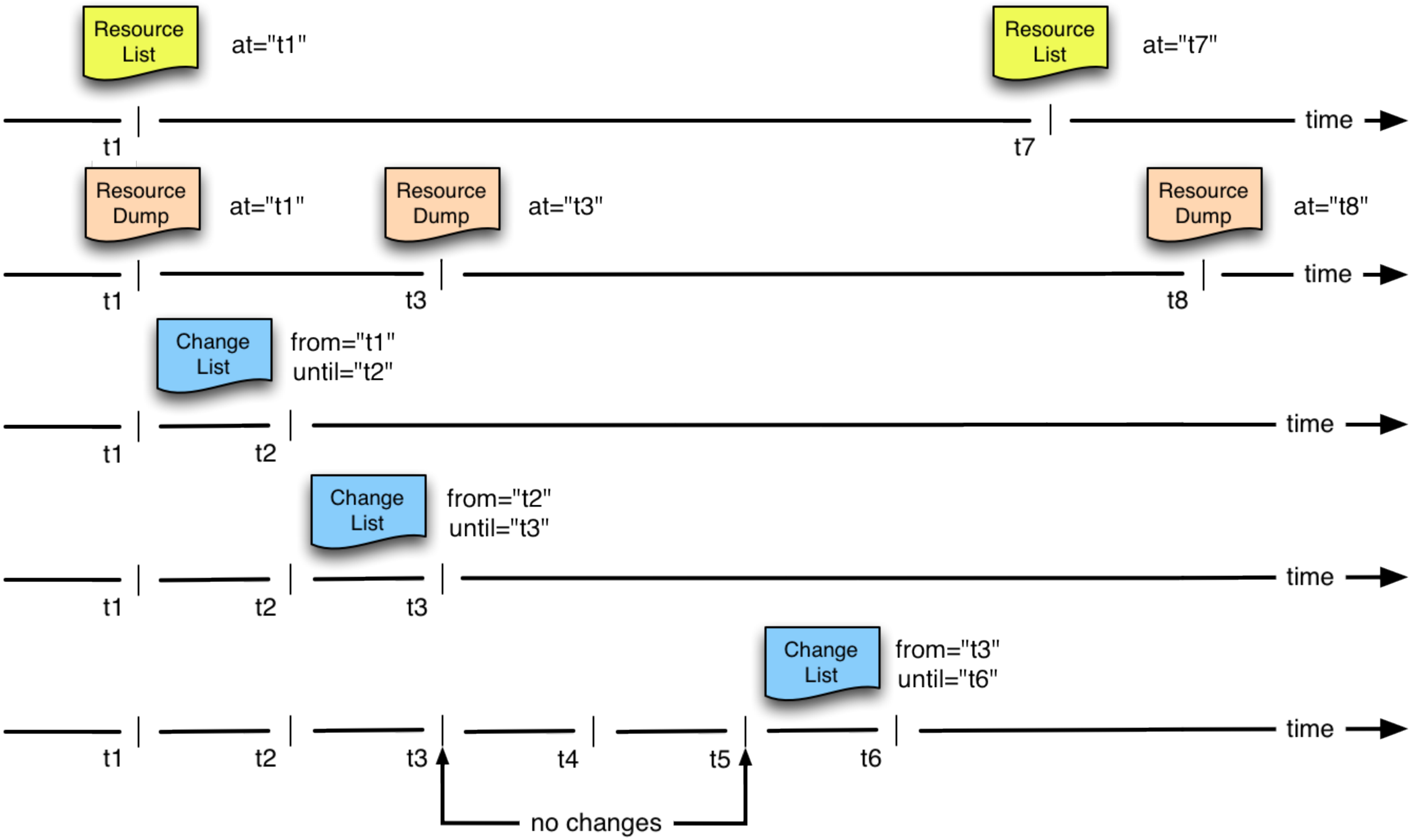
- URI
- Metadata
  - fixity
  - links

- URI
- Bitstream
- Metadata
  - fixity
  - links

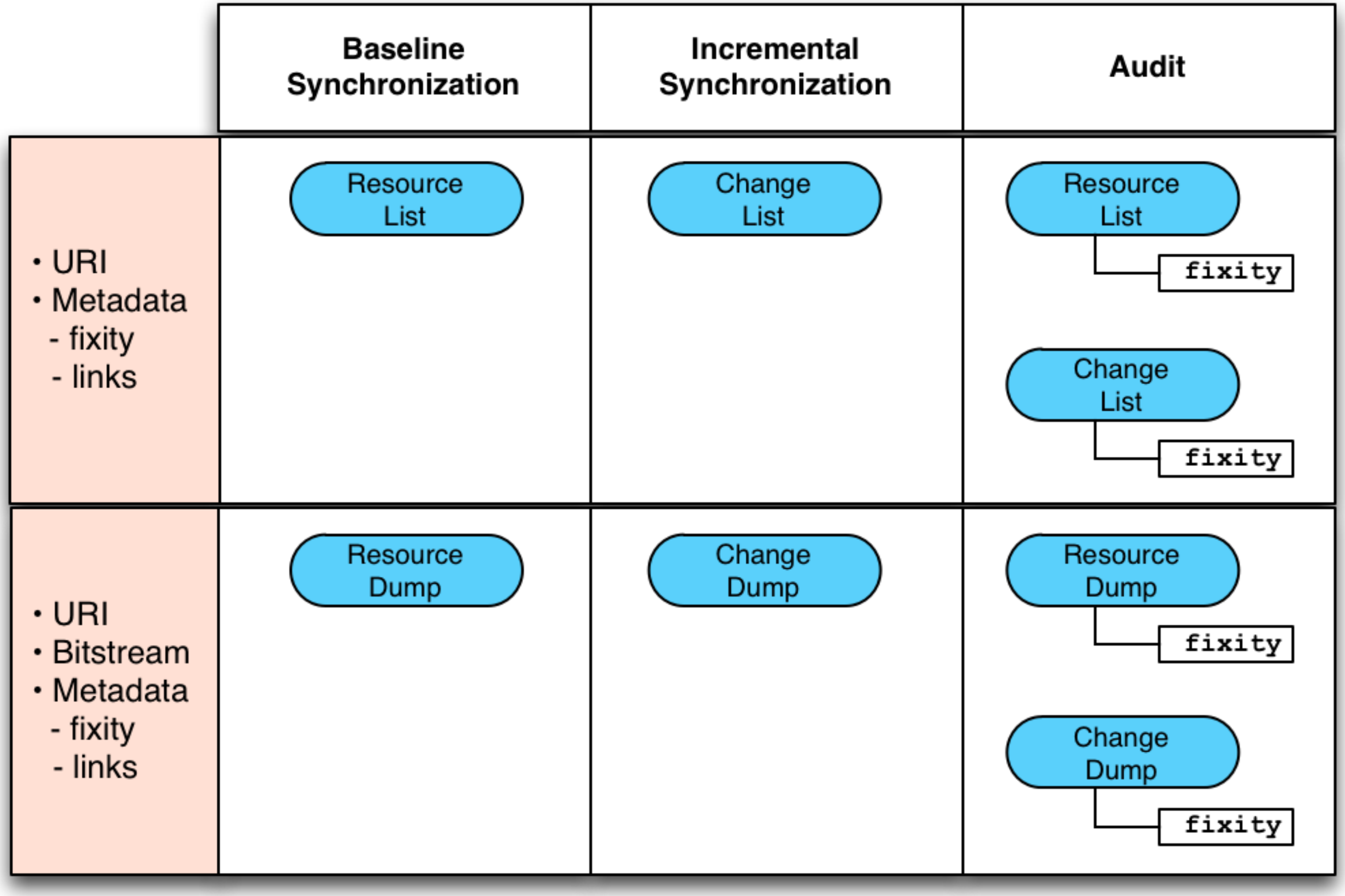
Source Capability 3: Describing Changes

In order to achieve lower latency and/or greater efficiency, a source may communicate about changes to its resources:

- Publish a **Change List**, a list of recent change events (created, updated, deleted resource)
  - Destination acts upon change events, e.g. GETs created/updated resources, removes deleted resources.
- A **Change List** pertains to resources that changed in a temporal interval with a start- and an end-date
  - If a resource changed more than once, it will be listed more than once


Resource List
at="t1"
Resource List
at="t7"
t1
t7
time
Resource Dump
at="t1"
Resource Dump
at="t3"
Resource Dump
at="t8"
t1
t3
t8
time
Change List
from="t1"
until="t2"
t1
t2
time
Change List
from="t2"
until="t3"
t1
t2
t3
time
Change List
from="t3"
until="t6"
t1
t2
t3
t4
t5
t6
time
no changes

Destination: Key Processes



	Baseline Synchronization	Incremental Synchronization	Audit
• URI • Metadata - fixity - links	Resource List	Change List	Resource List fixity Change List fixity
• URI • Bitstream • Metadata - fixity - links	Resource Dump	Change Dump	Resource Dump fixity Change Dump fixity

Menu

1. A personal spin
2. ResourceSync
   a. ResourceSync: Problem Perspective & Conceptual Approach
   b. Motivation & Use Cases
   c. Framework Walkthrough
   d. Framework Technical Details
   e. Implementation
3. ResourceSync and the Semantic Web

Many technology options

Push

DSNotify

Pull

OAI-PMH

rsync

Crawl

OAI-ORE

RDFsync

WebDAV Col. Syn.

XMPP

Atom

SWORD

AtomPub

Sitemap

RSS

PubSubHubbub

SPARQLpush

SDShare

XMPP

Many technology options

DSNotify

OAI-PMH

rsync

Crawl

RDFsync

Push

Pull

OAI-ORE

WebDAV Col. Syn.

XMPP

Atom

Sitemap

RSS

SWORD

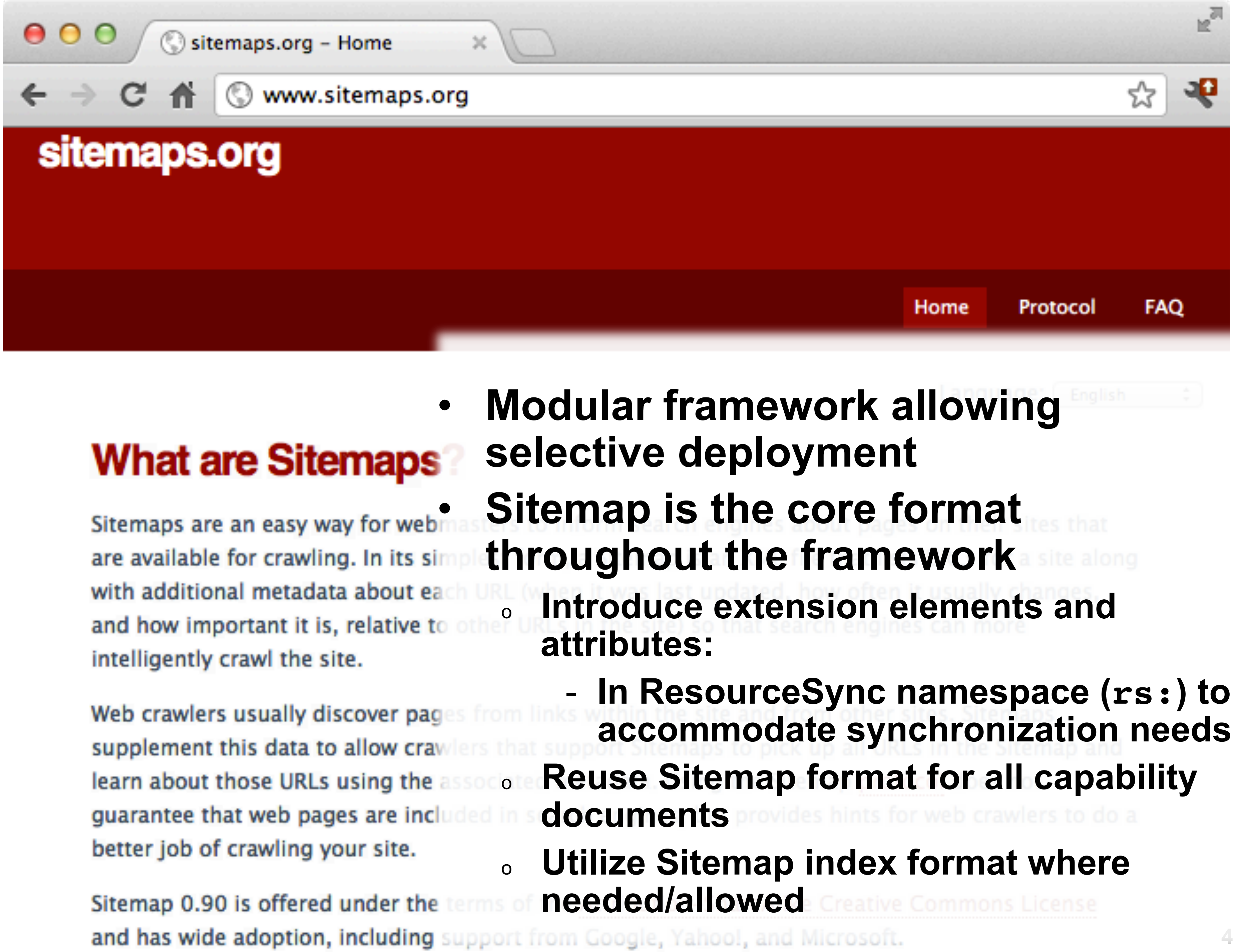
AtomPub

SPARQLpush

SDShare

XMPP

PubSubHubbub



What are Sitemaps?

Sitemaps are an easy way for webmasters to inform search engines about pages on their sites that are available for crawling. In its simplest form, a Sitemap is an XML file that lists URLs for a site along with additional metadata about each URL (when it was last updated, how often it usually changes, and how important it is, relative to other URLs in the site) so that search engines can more intelligently crawl the site.

Web crawlers usually discover pages from links within the site and from other sites. Sitemaps supplement this data to allow crawlers that support Sitemaps to pick up all URLs in the Sitemap and learn about those URLs using the associated metadata. Using the Sitemap protocol does not guarantee that web pages are included in search engines, but provides hints for web crawlers to do a better job of crawling your site.

Sitemap 0.90 is offered under the terms of the Attribution-ShareAlike Creative Commons License and has wide adoption, including support from Google, Yahoo!, and Microsoft.

- **Modular framework allowing selective deployment**
- **Sitemap is the core format throughout the framework**
  - **Introduce extension elements and attributes:**
    - **In ResourceSync namespace (`rs:`) to accommodate synchronization needs**
  - **Reuse Sitemap format for all capability documents**
  - **Utilize Sitemap index format where needed/allowed**

Sitemap Format

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9”>

  <url>
    <loc>http://example.com/res1</loc>
    <lastmod>2013-01-02T13:00:00Z</lastmod>
  </url>

  <url>
    <loc>http://example.com/res2</loc>
    <lastmod>2013-01-02T14:00:00Z</lastmod>
  </url>
  …
</urlset>
```

Use <sitemapindex> if >50k items

ResourceSync Sitemap Extensions

```
<urlset xmlns=http://www.sitemaps.org/schemas/sitemap/0.9
          xmlns:rs="http://www.openarchives.org/rs/terms/”>
  <rs:ln …/>
  <rs:md …/>

  <url>
    <loc>http://example.com/res1</loc>
    <lastmod>2013-01-02T13:00:00Z</lastmod>
    <rs:ln …/>
    <rs:md …/>
  </url>
  <url>
  …
  </url>
</urlset>
```

Same extensions in <sitemapindex>

Resource Metadata Summary

Element/Attribute	Description	Defined by
<loc>	Resource URI (identity)	sitemaps
<lastmod>	Timestamp of last change	sitemaps
<changefreq>	Expected update frequency	sitemaps
<rs:md>		ResourceSync
change	Change type (Change List & Change Dump Manifest only)	ResourceSync
encoding	HTTP Content-Encoding header value	RFC2616
hash	One or more content digests (md5, sha-1, sha-256)	Atom Link Ext.
length	HTTP Content-Length header value	RFC4287
path	Path in ZIP package (Dump Manifests only)	ResourceSync
type	HTTP Content-Type header value	RFC4287

Link Relation Summary

Relation	Use in ResourceSync	Defined in
rel="alternate"	Link from generic to specific URI	HTML 5
rel="canonical"	Link from specific to generic URI	RFC6596
rel="collection"	Resource is member of collection	RFC6573
rel="contents"	Link from dump to manifest	HTML4
rel="describedby"	Has metadata	Protocol for Web Description Resources (POWDER): Description Resources
rel="describes"	Is metadata for	The 'describes' Link Relation Type
rel="duplicate"	Mirror or alternative copy	RFC6249
rel=".../rs/terms/patch"	A patch -- efficient change information	This specification
rel="memento"	Link to time-specific URI	Memento Internet Draft
rel="timegate"	Link to timegate	Memento Internet Draft
rel="via"	Provenance chain, came from	RFC4287

ResourceSync Sitemap Validation

- All ResourceSync capability documents are valid according to the Sitemap XML Schema
  - http://www.sitemaps.org/schemas/sitemap/0.9

- For a more thorough validation use the ResourceSync XML Schema
  - http://www.openarchives.org/rs/0.9.1/resourcesync.xsd

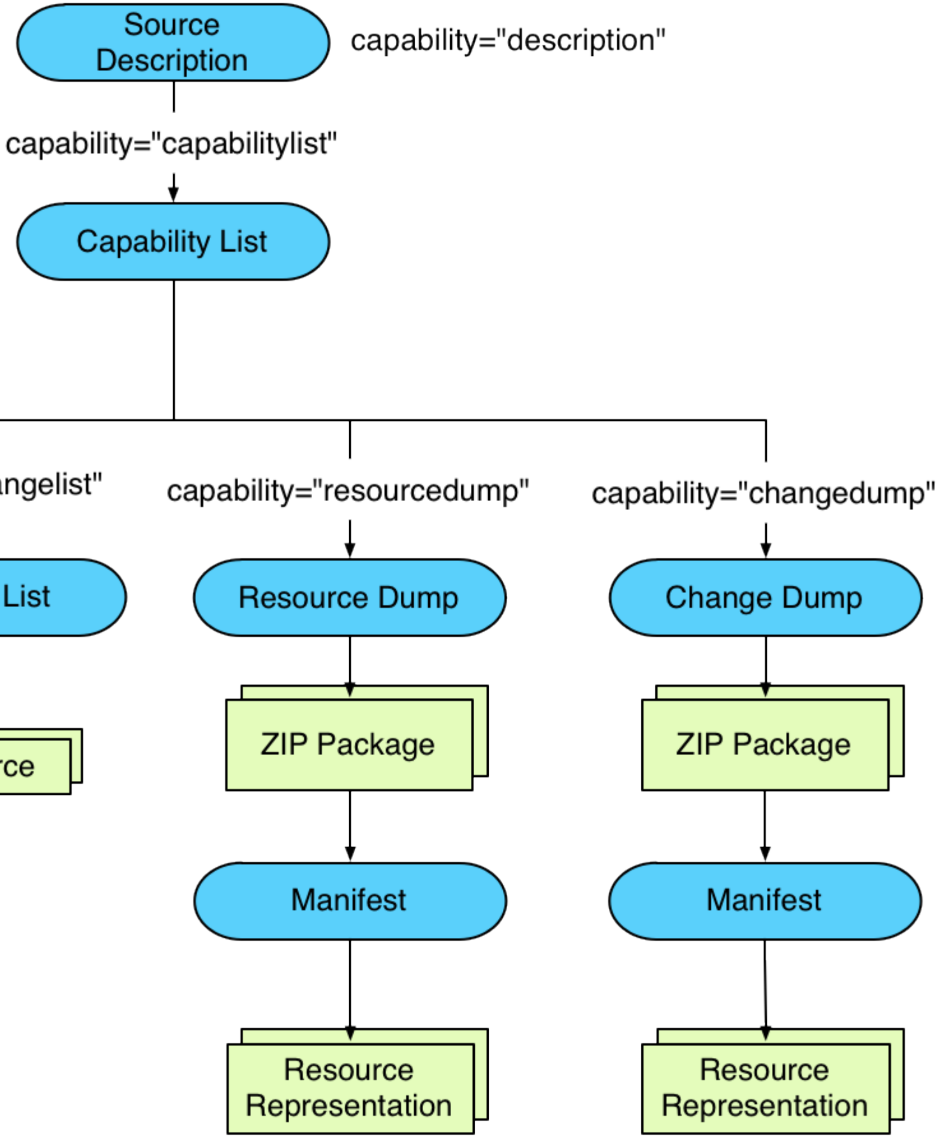
Example document: Resource List

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:rs="http://www.openarchives.org/rs/terms/">
  <rs:md capability="resourcelist"
         at="2013-01-03T09:00:00Z”
         completed="2013-01-03T09:01:00Z” />
  <url>
    <loc>http://example.com/res1</loc>
    <lastmod>2013-01-02T13:00:00Z</lastmod>
    <rs:md hash="md5:1584abdf8ebdc9802ac0c6a7402c03b6"
           length="8876"
           type="text/html"/>
  </url>
  <url>
  …
  </url>
</urlset>
```

- Describe Source’s resources subject to synchronization
  - At one point in time (snapshot)
  - Creation can take some time – duration can be conveyed
- HTTP GET resources

Framework Structure

(without possible index documents)



http://www.openarchives.org/rs/resourcesync#Structure

Supported Linking Use Cases

Provide links to related resources to address specific needs:

1. Mirrored content with multiple download locations
2. **Alternate representations of the same content**
   - Resources subject to HTTP content negotiation
   - Format migration for preservation reasons
3. Patching content rather than replacing it
4. Resources and metadata about resources
5. Prior versions of resources
6. Collection membership of resources
7. Republishing synchronized resources

All cases use <rs:ln> element referring to the linked resource

Linking – Alternate Representations – Case 1

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:rs="http://www.openarchives.org/rs/terms/">
  <rs:md capability=”changelist"
          from="2013-01-02T09:00:00Z”
          until="2013-01-03T09:00:00Z”/>
  <url>
    <loc>http://example.com/res1</loc>
    <lastmod>2013-01-02T13:00:00Z</lastmod>
    <rs:md change=”updated”/>
    <rs:ln rel="alternate"
           type="text/html"
           href="http://example.com/res1.html"/>
    <rs:ln rel="alternate"
           type=“application/pdf"
           href=”http://example.com/res1.pdf"/>
  </url>
</urlset>
```

Canonical URI links to specific URIs

Linking – Alternate Representations – Case 2

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
         xmlns:rs="http://www.openarchives.org/rs/terms/">
  <rs:md capability=”changelist"
           from="2013-01-02T09:00:00Z”
           until="2013-01-03T09:00:00Z”/>
  <url>
    <loc>http://example.com/res1.html</loc>
    <lastmod>2013-01-02T13:00:00Z</lastmod>
    <rs:md change=”updated”/>
    <rs:ln rel=”canonical”
             href="http://example.com/res1"/>
  </url>
</urlset>
```

Specific URI links to canonical URI

Source: Notification Capabilities

Motivations: reduce synchronization latency, avoid polling

P
U
S
H

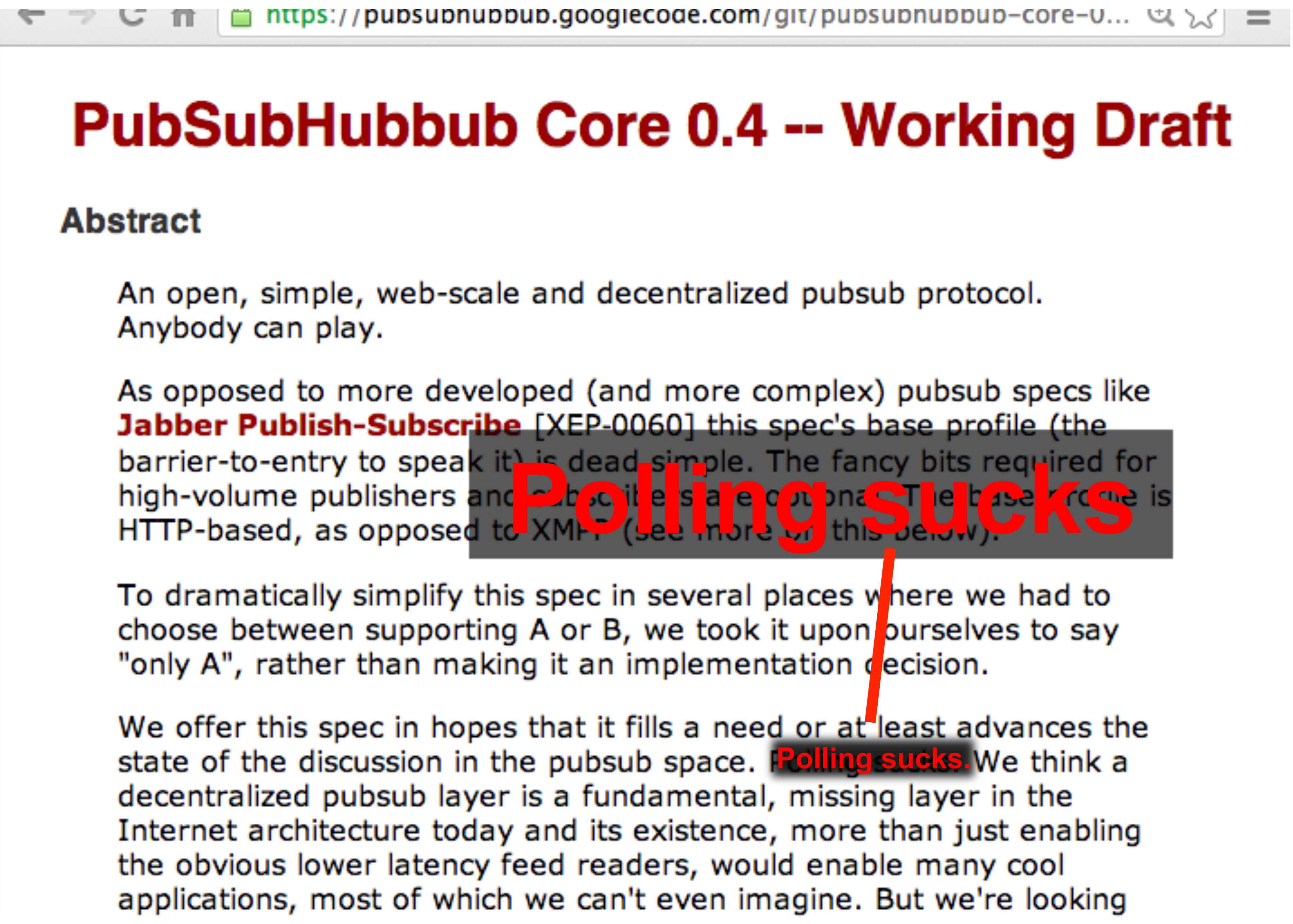
1. Change Notification

- Notifies destination about changes to particular resources
- e.g., resource A has been updated | created | deleted

2. Framework Notification

- Notifies destination about changes to capabilities i.e., their documents
- e.g., a Change List has been updated | created | deleted
- Also for Capability Lists and Source Description

Investigating Pubsubhubbub as transport first, may look at WebSockets later



https://pubsubhubbub.googlecode.com/git/pubsubhubbub-core-0...

PubSubHubbub Core 0.4 -- Working Draft

Abstract

An open, simple, web-scale and decentralized pubsub protocol. Anybody can play.

As opposed to more developed (and more complex) pubsub specs like **Jabber Publish-Subscribe** [XEP-0060] this spec's base profile (the barrier-to-entry to speak it) is dead simple. The fancy bits required for high-volume publishers and subscribers are optional. The base profile is HTTP-based, as opposed to XMPP (see more on this below).

To dramatically simplify this spec in several places where we had to choose between supporting A or B, we took it upon ourselves to say "only A", rather than making it an implementation decision.

We offer this spec in hopes that it fills a need or at least advances the state of the discussion in the pubsub space. Polling sucks. We think a decentralized pubsub layer is a fundamental, missing layer in the Internet architecture today and its existence, more than just enabling the obvious lower latency feed readers, would enable many cool applications, most of which we can't even imagine. But we're looking

Polling sucks

Source: Archival Capabilities

ARCHIVES

The Source may hold on to historical data, for example, to allow Destinations to catch up with events they missed or revisit prior resource states. To this end, the Source can publish archives, i.e. documents that enumerate historical capability documents

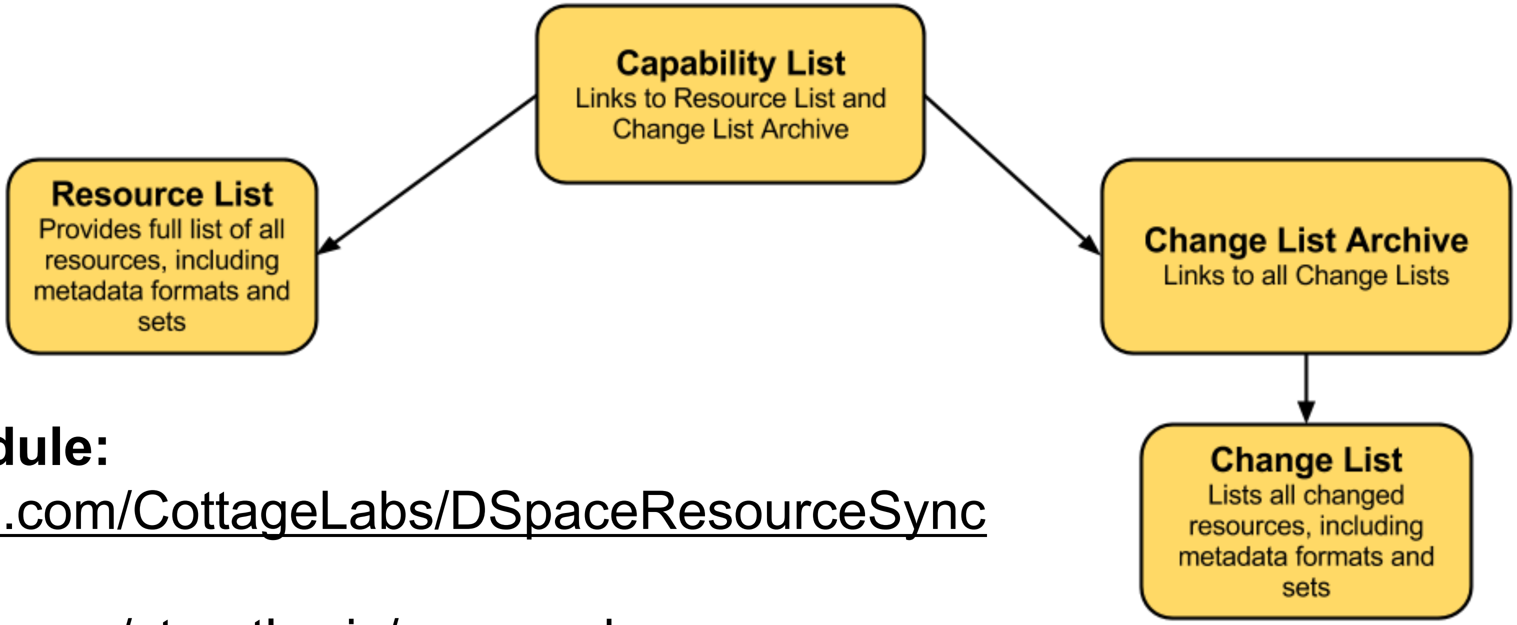
1. Resource List Archive
2. Resource Dump Archive
3. Change List Archive
4. Change Dump Archive

Re-use same document formats to list archived sets of corresponding documents, discovery entries tie together

Menu

1. A personal spin
2. ResourceSync
   a. ResourceSync: Problem Perspective & Conceptual Approach
   b. Motivation & Use Cases
   c. Framework Walkthrough
   d. Framework Technical Details
   e. Implementation
3. ResourceSync and the Semantic Web

DSpace support for metadata harvesting use case



Capability List
Links to Resource List and Change List Archive

Resource List
Provides full list of all resources, including metadata formats and sets

Change List Archive
Links to all Change Lists

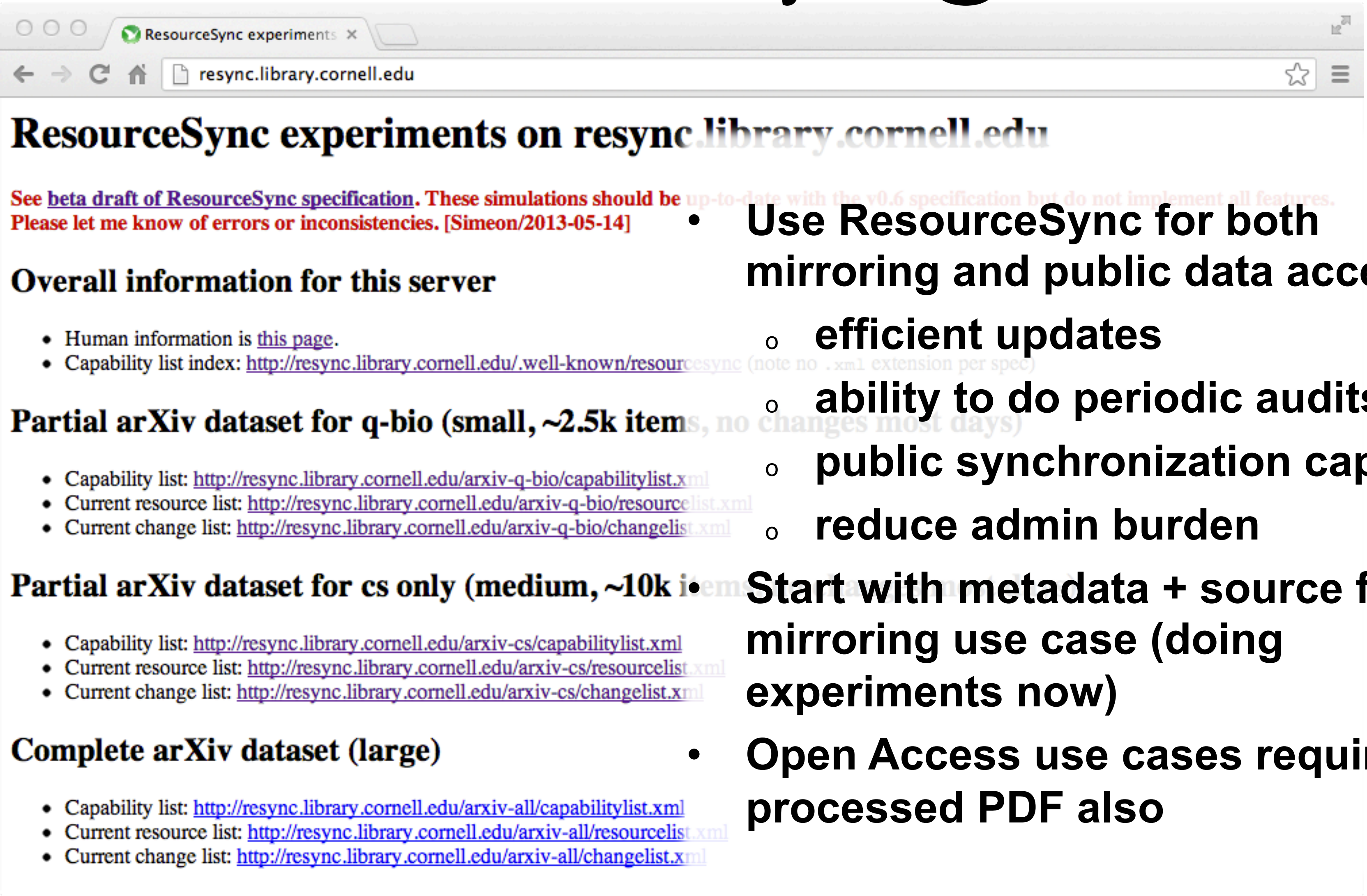
Change List
Lists all changed resources, including metadata formats and sets

DSpace Module:
https://github.com/CottageLabs/DSpaceResourceSync
PHP client:
https://github.com/stuartlewis/resync-php

```
http://mydspace.edu/dspace-rs/resource/123456789/7/qdc
```

ResourceSync webapp **Item handle** **Metadata Format**

ResourceSync @ arXiv



ResourceSync experiments on resync.library.cornell.edu

See beta draft of ResourceSync specification. These simulations should be up-to-date with the v0.6 specification but do not implement all features. Please let me know of errors or inconsistencies. [Simeon/2013-05-14]

Overall information for this server

- Human information is this page.
- Capability list index: http://resync.library.cornell.edu/.well-known/resourcesync (note no .xml extension per spec)

Partial arXiv dataset for q-bio (small, ~2.5k items, no changes most days)

- Capability list: http://resync.library.cornell.edu/arxiv-q-bio/capabilitylist.xml
- Current resource list: http://resync.library.cornell.edu/arxiv-q-bio/resourcelist.xml
- Current change list: http://resync.library.cornell.edu/arxiv-q-bio/changelist.xml

Partial arXiv dataset for cs only (medium, ~10k items, changes most days)

- Capability list: http://resync.library.cornell.edu/arxiv-cs/capabilitylist.xml
- Current resource list: http://resync.library.cornell.edu/arxiv-cs/resourcelist.xml
- Current change list: http://resync.library.cornell.edu/arxiv-cs/changelist.xml

Complete arXiv dataset (large)

- Capability list: http://resync.library.cornell.edu/arxiv-all/capabilitylist.xml
- Current resource list: http://resync.library.cornell.edu/arxiv-all/resourcelist.xml
- Current change list: http://resync.library.cornell.edu/arxiv-all/changelist.xml

- **Use ResourceSync for both mirroring and public data access**
  - **efficient updates**
  - **ability to do periodic audits**
  - **public synchronization capability**
  - **reduce admin burden**
- **Start with metadata + source for mirroring use case (doing experiments now)**
- **Open Access use cases require processed PDF also**

Python Library and Client

- Aim to provide library code implementing all ResourceSync facilities for use in both source and destination implementations
  - Designed for python 2.6 (RHEL6) and 2.7
  - Will not work with python <= 2.5
- Client (`resync`) supports many destination operations, inspired by the common Unix `rsync` program
- Client also supports some operations that might be useful in a source, such as generation of static Resource Lists, or periodic Change Lists (used in arXiv experiments)
- Explorer (`resync-explorer`) intended to allow easy inspection of a source's resource sets and capabilities
- Developed since ResourceSync v0.5, updated for v0.9.1

http://github.org/resync/resync

ResourceSync Source Simulator

- Python code using Tornado server
- Provides random set of resources of different sizes updated at a particular rate
- Very useful for testing Destination code

http://github.com/resync/simulator

Menu

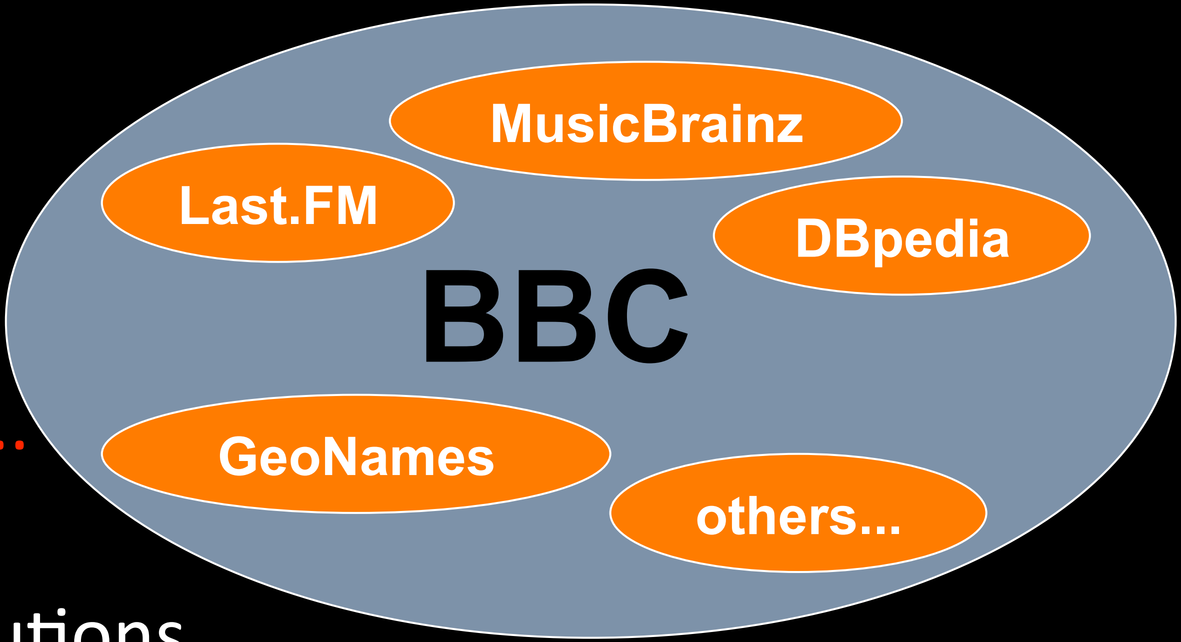
1. A personal spin
2. ResourceSync
   a. ResourceSync: Problem Perspective & Conceptual Approach
   b. Motivation & Use Cases
   c. Framework Walkthrough
   d. Framework Technical Details
   e. Implementation
3. ResourceSync and the Semantic Web

Linked data

Fundamentally distributed but local copy often required. Either:

1. cache

2. sync local copy...



- Many ad-hoc solutions for local copy

How do you get your semantic data?

- @edsu -> @LibSkrat: “not at http://id.loc.gov no; someone could download the triples and create their own though”
- Philipp Zumstein – “copy the RDF/XML files”
- Valeria Pesce – “harvest XML and CSV, then map to extended VIVO ontology”
- ...

- Poll on storage: triple store? files? RDB? other?

Semantic data synchronization

- Is your data nice linked data? URIs that resolve to other documents, etc.
  - everything is a web resource so good match for ResourceSync
  - maybe the web already provides adequate access?
- Are you able to tell what has changed?
  - in most triple stores there is no timestamp so providing subsets of changed data might be hard
- Look at four scenarios...

Linked data

Sematic data **on the web** – great match for ResourceSync

Consider a linked data system has some convenient way to generate and/or keep track of fixity information (datestamps, hashes, etc.) for all of its resource representations, then this may be an effective way to synchronize with ResourceSync.

- Usual ResourceSync mechanisms including Resource Lists, Change Lists, Dumps, Notifications and Archives all applicable.
- Complications with triple store
  - How to generate fixity information?
  - Cost of generating set of self-contained representations (e.g. concise bounded description) may be high

Service level notification

Image a set of RDF data updated periodically, goal is just to let consumers know that changes have been made

- Perhaps many sets or subsets provided
- Need to give service a URI which is then listed in Resource List etc.

Dumps as resources to sync

- Might have dump that is in any format, it is just a resource on the web
- Almost trivial but fits very cleanly in framework
- Would work well with framework notification

- Use link relation might also be used to indicate sequence if there are a set of dumps from different times

Note: Somewhat different from a ResourceSync Resource Dump (or Change Dump) which is something where the data is represented as resources to be synchronized

Diffs or patches for RDF data

Open question: it might be possible to use RDF patching mechanisms (perhaps JSON-PATCH with JSON-LD) to provide efficient updates of RDF datasets

- Trivial for a dataset with no blank nodes,
- Diffs progressively more difficult and less efficient if there are many blank nodes
- Particularly useful/efficient for large datasets with small changes

ResourceSync provides mechanism to link any patch format and file, and relate to the resource patched

That’s all folks

Pointers

- **Specification**

  `http://www.openarchives.org/rs/`

  `http://www.openarchives.org/rs/resourcesync`

  `http://www.openarchives.org/rs/archives`

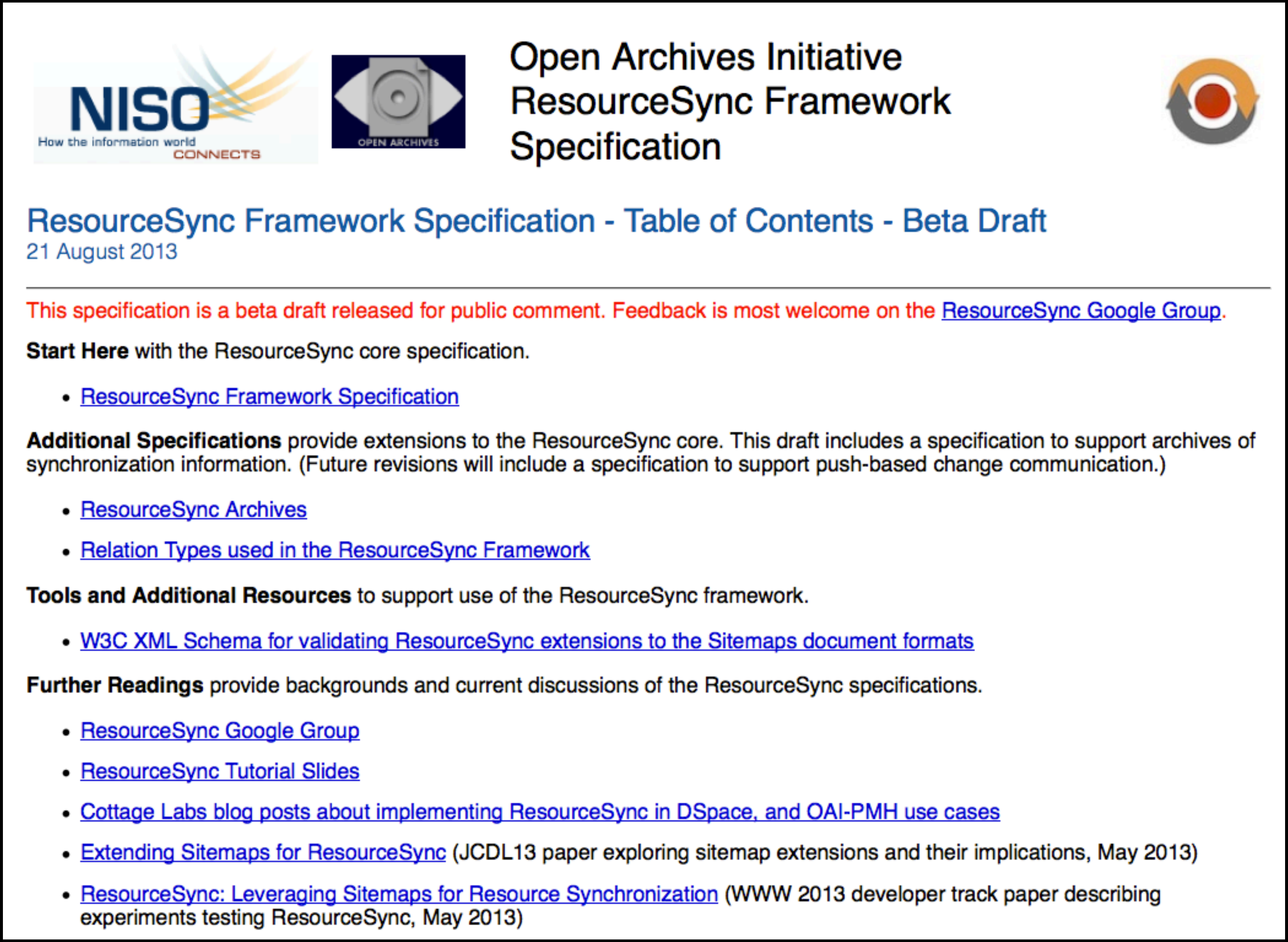
- **List for public comment**

  `https://groups.google.com/d/forum/resourcesync`

- **Client and simulator code**

  `http://github.org/resync/resync`

  `http://github.org/resync/simulator`



NISO How the information world CONNECTS

OPEN ARCHIVES

Open Archives Initiative ResourceSync Framework Specification

ResourceSync Framework Specification - Table of Contents - Beta Draft

21 August 2013

This specification is a beta draft released for public comment. Feedback is most welcome on the ResourceSync Google Group.

Start Here with the ResourceSync core specification.

- ResourceSync Framework Specification

Additional Specifications provide extensions to the ResourceSync core. This draft includes a specification to support archives of synchronization information. (Future revisions will include a specification to support push-based change communication.)

- ResourceSync Archives
- Relation Types used in the ResourceSync Framework

Tools and Additional Resources to support use of the ResourceSync framework.

- W3C XML Schema for validating ResourceSync extensions to the Sitemaps document formats

Further Readings provide backgrounds and current discussions of the ResourceSync specifications.

- ResourceSync Google Group
- ResourceSync Tutorial Slides
- Cottage Labs blog posts about implementing ResourceSync in DSpace, and OAI-PMH use cases
- Extending Sitemaps for ResourceSync (JCDL13 paper exploring sitemap extensions and their implications, May 2013)
- ResourceSync: Leveraging Sitemaps for Resource Synchronization (WWW 2013 developer track paper describing experiments testing ResourceSync, May 2013)