ResourceSync – An Introduction

Todd Carpenter
Executive Director, NISO
Wolfram Data Summit
Thursday, September 6, 2012

With thanks to Herbert Van de Sompel and Robert Sanderson (LANL)
@TAC_NISO Twitter Highlights

- Presenting this afternoon on the ResourceSync project at Wolfram Data Summit #wolframsummit
- I’m pre-tweeing my slides during #rsync presentation. Slides will be posted later today #wolframsummit
- NISO mission develop & maintain technical standards related to information, documentation, discovery & distribution of content #wolframsummit
- Machines don’t talk like people do. Then again some people don’t talk like other people do, particularly teenagers #wolframsummit
- So where did the ResourceSync project start? #NISO approached OAI about updating the PMH protocol. #wolframsummit
- The #NISO / OAI ResourceSync project was possible through the generous support of the Alfred P. Sloan Foundation. Thank you! #wolframsummit
- What is RSync trying to solve (1/2): Source Server has resources that change. Destination servers want to leverage some/all of Source #wolframsummit
- What is RSync trying to solve (2/2): How to sync on ongoing basis in near-real-time & at web scale with as little system overhead as poss #wolframsummit
- RSync studied # of existing protocols to determine protocols that best meet needs. Bias against developing new spec from scratch. #wolframsummit
- The goal of ResourceSync is to find the model that most efficiently distributes the content, while limiting the tax on the source system. #wolframsummit
- Very early days in process of standards development. Still in incubation stage. Consensus & adoption phases coming ’13 & beyond #wolframsummit
- Draft alpha specification of ResourceSync posted in August, Team meeting in Sept to review comments #wolframsummit
About

Non-profit industry trade association accredited by ANSI

Mission of developing and maintaining technical standards related to information, documentation, discovery and distribution of published materials and media

125+ Member organizations

Volunteer driven organization: 400+ spread out across the world

Represent US interests to ISO in the areas of Information & Documentation
Standards are familiar, even if you don’t notice
Machines don’t talk like people do

Sometimes I wish you had a little green light that would blink on and off like the computer, so I could tell whether you’re processing anything or if you’ve just frozen up.
Machines talk like this
How did we get here?

• OAI–PMH Protocol
  – Developed in 2001 (v 1.1, v 2.0 – 2002)
  – Developed by Herbert van de Sompel, Carl Lagoze, Michael Nelson, and Simeon Warner
  – Fairly wide adoption in scholarly community

• In spring 2011, NISO approached OAI to discuss updating PMH Protocol

• Response was “Let’s try something else more in line with more modern
A partnership is born
A partnership is born

Agreement to launch RSync as a NISO standards initiative
A partnership is born

Agreement to launch RSync as a NISO standards initiative

Partnership on grant application
A partnership is born

Agreement to launch RSync as a NISO standards initiative

Partnership on grant application

OAI team comprised core technology team
A partnership is born

Agreement to launch RSync as a NISO standards initiative

Partnership on grant application

OAI team comprised core technology team

Vetting & education by NISO
Special thanks are due to...
ResourceSync Working Group

**Herbert Van de Sompel (Chair)**
Los Alamos National Laboratory

**Todd Carpenter (Co-Chair)**
National Information Standards Organization (NISO)

**Nettie Lagace**
National Information Standards Organization (NISO)

**Manuel Bernhardt**
Delving B.V.

**Kevin Ford**
Library of Congress

**Bernhard Haslhofer**
Cornell University

**Richard Jones**
Joint Information Systems Committee (JISC)

**Martin Klein**
Los Alamos National Laboratory

**Graham Klyne**
Joint Information Systems Committee (JISC)

**Carl Lagoze**
Cornell University

**Stuart Lewis**
Joint Information Systems Committee (JISC)

**Peter Murray**
Lyrasis

**Michael Nelson**
Old Dominion University
David Rosenthal
Stanford University

**David Rosenthal**
LOCKSS

**Christian Sadilek**
Red Hat

**Shlomo Sanders**
Ex Libris, Inc.

**Robert Sanderson**
Los Alamos National Laboratory

**Sjoerd Siebinga**
Delving B.V.

**Ed Summers**
Library of Congress

**Simeon Warner**
Cornell University

**Jeff Young**
OCLC Online Computer Library Center
What are we trying to do?

• Synchronize web resources – things with a URI that can be dereferenced and are cache-able
• Improve on web synchronization methods
• For small websites/repositories (a few resources) to large repositories/datasets/linked data collections (many millions of resources)
• That change slowly or rapidly
• Focus on needs of research and cultural heritage organizations, but aim for generality
Use Cases

One-to-one sync

Master Copy (one to many)

Aggregator (many to one)

Selective sync
More Use Cases

XML metadata harvesting

Protected Resources

Large data files

Statistics collection

Protected Resources

Protected Resources

Protected Resources

Protected Resources
Not (yet) Use Cases
(i.e.: Out of Scope)

• Bidirectional synchronization

• Destination-defined selective synchronization (query)

• Bulk URI migration
Use cases differ

How good is the synchronization?

Perfect <-> Good enough

How fast is the synchronization?

Fast <-> Fast enough
3 distinct needs regarding resource synchronization

Baseline matching: An approach to allow a Destination that wants to start synchronizing with a Source to perform an initial catch up – Dump.

Incremental resource synchronization: An approach to allow a Destination to remain up-to-date regarding changes at the Source.

Audit: An approach to allow checking whether a Destination is in sync with a Source – Inventory.

=> All 3 are considered in scope for ResourceSync
Incremental Synchronization

Change Notification (CN)
Alert that something happened
(create, update, delete)

Content Transfer (CT)
Transfer of just the change or the full resource
Trivial versus Optimal

• Trivial Approach – Retrieve & Compare

• Optimal Approach – push only the change to only the destinations monitoring the resource
More advanced option

Feed Extension Solution:
Continue the Feed paradigm, but introduce aggregating service and ping notification to re-pull (simulated push)
Only advantageous if Source already supports a Feed
Change Notification – Protocols

- Atom PubSubHubbub (PuSH)
- XMPP
  - PubSub extension
  - BoSH (XMPP over HTTP)
- Comet / HTTP Streaming
  - Open an HTTP connection and keep reading from it
- Bayeux Protocol
- Long Polling
  - Keep HTTP connection open until a message, then reopen
- BoSH, Bayeux option
- WebSockets
  - NullMQ / ZeroMQ
- XMPP over WebSockets?
How standards proliferate:
(See: A/C chargers, character encodings, instant messaging, etc.)

Situation: There are 14 competing standards.

14? Ridiculous!
We need to develop one universal standard that covers everyone's use cases. Yeah!

Soon:

Situation: There are 15 competing standards.

http://imgs.xkcd.com/comics/standards.png/
RSync Alpha
A Framework Based on Sitemaps

• Developing a Modular framework allowing selective deployment
• Sitemap is the core component throughout the framework
• Introducing extension elements and attributes:
  – In ResourceSync namespace (rs:) to accommodate synchronization needs
  – In XHTML namespace (xhtml:) mainly to accommodate discovery needs
• Reuse Sitemap format for Change Sets (both current and historical) and for manifest in Dump
Communications structure

Pushing Change Sets via XMPP PubSub

- Publish to foo.org/node1
- Subscribe to foo.org/node1
- XMPP server-server
- bar.org
- joe@foo.org
- jane@bar.org
The lifecycle of standards

![Standards Development Lifecycle Diagram](image-url)

You are here
Timeline

• Project Launch = November 2011
• Approved work item = December 2011
• Working Group formed = February 2012
• Webinar on project = March 2012
• JCDL meeting, Washington DC = June 2012
• Alpha = September 2012
• Team meeting, Denver, CO = September 2012
  – forthcoming D–Lib article
• Beta/Draft for trail use = ?? December 2012
• Comment period = ?? December 2012 – March 2012
• Training = ?? May – July 2013
• Approval = ?? December 2013
More information

Background webinar (March 6, 2012) recording

First draft spec: http://www.openarchives.org/rs/0.1/resourcesync

Simulator code on github http://github.org/resync/simulator

NISO workspace http://www.niso.org/workrooms/resourcesync/

List for public comment coming soon
Standards for Data/Exchange: New Work areas?

Many potential areas for work in sharing of data including:

• Author/Contributor disambiguation & other issues

• Data Equivalence – How does one know that this thing and that are equivalent (i.e., contain same data)?

• Systemic metadata
  What is the form of this information?
  What are its structural components?

• Archival issues
  Storage, physical level, metadata, but also migration issues

• Bibliographic information for discovery, delivery and reuse

• Bibliometrics / Assessment & impact

• Rights issues – Ownership, recognition, sharing, privacy
What are appropriate metrics?

- For datasets, what is a download?
- How does one use compare with another?
- Citation ecosystem needs to develop
- Data papers?
Data Equivalence

Basically, is an Excel file equivalent to a text file?

Creation of a high-level conceptual model of data description

A “FRBR” for data

Defines the distinctions between states & transformations of data

Basis for identification & description
Thank you!

Todd Carpenter
Executive Director
tcarpenter@niso.org

National Information Standards Organization (NISO)
3600 Clipper Mill Road, Suite 302
Baltimore, MD 21211 USA
+1 (301) 654-2512
www.niso.org

NOTE => NISO HAS MOVED!! <=