Establishing New Levels of Interoperability for Web-Based Scholarship

Herbert Van de Sompel
Los Alamos National Laboratory
@hvdsomp


Cartoon by:
Patrick Hochstenbach
Reminiscing About 15 Years of Interoperability Efforts

Herbert Van de Sompel
Los Alamos National Laboratory
herbertv@lanl.gov

Michael L. Nelson
Old Dominion University
mln@cs.odu.edu

DOI: 10.1045/november2015-vandesompel
Research Communication & Research Process on the Web

• A highly distributed activity

• Turning this distributed activity from a gathering of silo-ed nodes into an ecology of collaborating nodes, requires establishing interoperability
  • In the web context, this seems like a rather unique challenge: Most web enterprises do not want interoperability they want dominance, monopoly

• To a large extent, interoperability across this distributed activity remains restricted to persistent identification of communicated objects and contributors
  • Which results in added-value – services can be created
Evolution of Our Thinking about Interoperability

interoperability

- repository centric
  - OAI-PMH
- web centric
  - OAI-ORE
    - descriptive - RDF
    - navigational - HTTP Links

Memento
Evolution of Our Thinking about Interoperability

- interoperability
  - repository centric
  - web centric
    - descriptive - RDF
    - navigational - HTTP Links
• OAI was a heroic effort to fundamentally transform scholarly communication
  • By promoting communication via preprints, non-peer-reviewed papers

• The OAI took a technical approach to achieve the goal
  • Make preprints easier to discover, access

http://www.openarchives.org/OAI/openarchivesprotocol.html
The OAMH protocol is a low-barrier interoperability specification for the recurrent exchange of metadata between systems
Those Were the Days

Lycos™
Lycos search form, large catalog

Query: 
Max-hits: 15  Min-terms: 1  Min-score: 0.01  Terse output: 
Start search  Reset

AltaVista™
The most powerful and useful guide to the
Ask AltaVista™ a question. Or enter a few words in
any language

+ bats + animals

Search

Herbert Van de Sompel
COAR Annual Meeting, Vienna, Austria, 12/04/2016
http://
3.1.1.1 Encoding an OAI-PMH request in a URL for an HTTP GET

Don’t trust HTTP

3.6 Error and Exception Conditions

In event of an error or exception condition, repositories must indicate OAI-PMH errors, distinguished from HTTP Status-Codes, by including one or more error elements in the response. While one
HTTP GET with GetRecord verb

http://an.oa.org/OAI-script?
A repository replies to a request with an *incomplete list* and a resumptionToken;
Repository-Centric Interoperability Paradigm

Address interoperability challenges from the perspective of a node, e.g. an IR, a publisher, a web-based authoring portal, a software repository, …

- **The node at the center of the universe**

- Define a machine interface for your node, expect others to use it

- Piggybacking on the web without truly embracing its core technologies

- The node resembles a brick & mortar library that can be visited subject to well-intended yet idiosyncratic policies – the interface
Launching the OAI - Luce, Van de Sompel, Ginsparg (1999)

Repositories still use OAI-PMH, created in the olden days when I looked like this
Evolution of Our Thinking about Interoperability

interoperability

repository centric

web centric

descriptive - RDF

navigational - HTTP Links
Web-Centric, Resource-Centric Interoperability Paradigm

Address interoperability challenges from the perspective of the web

- The resource at the center of the universe
  - The notion of a node, a repository, not even of a web server exists in the architecture of the web

- The tools of the interoperability trade are the primitives of the web
Tools of the Web-Centric Interoperability Trade

- Resource
- URI
- HTTP as the API: HEAD/GET, POST, PUT, DELETE
- Representation
- Media Type
- Link
- Content Negotiation
Evolution of Our Thinking about Interoperability

- interoperability
  - repository centric
  - web centric
    - descriptive - RDF
    - navigational - HTTP Links
• OAI-ORE observation: Scholarly assets are rapidly becoming *compound*, consisting of multiple resources with various:
  • Relationships
  • Interdependencies

• How to convey this compound-ness in an interoperable manner so that applications can access, consume such assets?

http://www.openarchives.org/ore/1.0/toc
Aiming for New Levels of Cross-Repository Functionality
TICER, Digital Libraries a la Carte, Tilburg, The Netherlands, August 22 2006
Herbert Van de Sompel
ORE Will Allow Web Crawlers to Unambiguously Recover CDO Structure from the Web Graph
Express the `ore:describes` relationship
Tools of the Web-Centric Interoperability Trade – RDF Stack

- Resource
- URI
- HTTP as the API
- Representation
- Media Type
- Link
- Content Negotiation, e.g. for preferred Media Type

- Typed Link
- Controlled Vocabularies for Typed Links

W3C Architecture of the World Wide Web

RDF, RDFS, OWL
Interoperability via RDF, RDFS, OWL Stack

Used by various interoperability efforts, e.g. OAI-ORE, Open Annotation, W3C PROV, Research Objects, …

• Provides extensive expressiveness for description
• Typically based on publishing documents that adhere to a certain “profile” and reveal relations, properties, …
• Non-Trivial barrier to entry as illustrated by slow adoption, likely related to unfamiliar technology stack
Evolution of Our Thinking about Interoperability
• Memento is about the Web and time:
  • Resources evolve over time
  • Only the current resource version is available from a resource’s URI
  • How to seamlessly access prior versions, if they exist, using the resource’s URI and a version datetime

• Memento looks at this problem for the Web, in general:
  • Time-Based access to resource versions across web archives, resource versioning systems

RFC7089 - http://mementoweb.org/guide/rfc/
Original Resource and Mementos

Original Resource

URI-R

current representation

URI-M_1

Apr 10 2001 representation

URI-M_2

Aug 15 2007 representation

Mementos
Bridge from Present to Past
Bridge from Present to Past

URI-R \rightarrow \text{HTTP Link timegate} \rightarrow \text{URI-G}

\text{Datetime Negotiation Accept-Datetime} = T_j

\text{URI-M}_1 \rightarrow \text{Mementos}

\text{Aug 15 2007 representation}

\text{URI-M}_2

\text{URI-G}

\text{TimeGate}

\text{Original Resource}

current representation

Herbert Van de Sompel
COAR Annual Meeting, Vienna, Austria, 12/04/2016
Bridge from Past to Present

URI-R

HTTP Link original

current representation

Original Resource

URI-M₁

URI-M₂

HTTP Link original

Apr 10 2001 representation

Aug 15 2007 representation

Mementos
Tools of the Web-Centric Interoperability Trade – HTTP Stack

- Resource
- URI
- HTTP as the API
- Representation
- Media Types
- Link
- Content Negotiation, e.g. for Media Type, Time
- Typed Link
- Controlled Vocabularies for Typed Links

HATEOAS – Hypermedia As The Engine Of Application State

http://en.wikipedia.org/wiki/HATEOAS
Interoperability via HTTP Links, IANA Link Relation Types

Used by Memento, ResourceSync, Signposting the Scholarly Web:

- Provides **coarse expressiveness for navigation** via IANA registered relation types (expressed as reserved terms)
  - Finer grained expressiveness via community-defined relation types (expressed as HTTP URIs)
- Typically based on publishing typed links that support a client to navigate among resources in an informed manner
- Low implementation barrier because of familiar technology stack
Establishing New Levels of Interoperability: Examples

ResourceSync

Signposting the Scholarly Web

Robust Links
Anurag Acharya Told Us Why We Need ResourceSync

What does indexing need?

- List of all article urls
- Ability to fetch article urls
- What we index is what the user sees

- Identify scholarly articles
- Determine article metadata
Anurag Acharya Told Us Why We Need ResourceSync

List of articles - IV

- Best practice: Year-month browse
  - Linked from homepage - EPrints
  - Helps crawlers as well as users
- Best practice: Article sitemap
  - Include urls for ALL articles
  - Linked from robots.txt or homepage
  - DSpace if sitemaps are enabled

ResourceSync is Based on Sitemaps

- Sitemap is the document format used throughout the framework
  - Used widely by web servers to advertise their resources to search engines

```xml
<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>
```

- Synchronization of resources from a Source to Destinations
  - Includes exposing repository content to aggregators, search engines
- Applies to any resource with an HTTP URI
- Leverages key ingredients of web interoperability, follow your nose, existing Search Engine Optimization practice

http://www.openarchives.org/rs/toc
Publish Inventory, Changes, Notifications

• Repository communicates about the state of its resources:
  
  • **Publish inventory**: snapshot of the state of resources at a moment in time
  
  • **Publish changes**: enumeration of resource changes that occurred during a temporal interval
  
  • **Notify about changes**: send notifications as changes occur
Payload for Inventory, Changes, Notifications

- A repository may communicate additional information pertaining to each resource:
  - **Technical metadata about a resource**: content encoding, content length, mime type, content-based hash
  - **Links to related resources**: mirror copies, alternate representations, resource versions, diff between current and previous version, metadata-to-content link, content-to-metadata link, collection membership, persistent identifier, etc.
ResourceSync is Based on Sitemaps

- Extensions to Sitemaps:
  - `<rs:ln>` for links
  - `<rs:md>` for metadata

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

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

...<urlset>
```
Signposting the Scholarly Web

Example pattern: The PID, the Landing Page, the Stuff
Response to HTTP HEAD on http://dx.doi.org/10.2218/ijdc.v9i1.320

HTTP/1.1 303 See Other
Server: Apache-Coyote/1.1
Date: Fri, 9 Jan 2015 16:31:46 GMT
Vary: Accept
Location: http://www.ijdc.net/index.php/ijdc/article/view/320
Link: <http://www.ijdc.net/index.php/ijdc/article/view/320>; rel=”describedby”
; type=“text/html”
Content-Length: 188
This Allows a Machine Agent …

• To understand that the splash page describes the DOI-identified asset

• To determine that resource A is not part of the DOI-identified asset

• To navigate towards the profile of the authors of the asset when landing on any of the constituent resources of the DOI-identified asset

• To understand that a DOI is associated with the PDF, HTML, and JPEG resources and that this DOI should preferably be used to refer to those resources

• To associate annotations made to the HTML page with the DOI
Signposting: Work in Progress

Signposting the Scholarly Web
Signposting aims to achieve meaningful interoperability for web-based scholarship.

Signposting is a collaboration between the Prototyping Team of the Research Library of the Los Alamos National Laboratory and the Computer Science Department of Old Dominion University.

Demo
Input any HTTP URI of a scholarly article, and hit Get Headers to see its corresponding signposting headers.

An HTTP URI of a scholarly article.
Signposting: Work in Progress

Demo
Input any HTTP URI of a scholarly article, and hit Get Headers to see its corresponding signposting headers.

DOI: http://dx.doi.org/10.1371/journal.pone.0115253

Signposting Headers for the Landing Page Pattern

<table>
<thead>
<tr>
<th>Resource</th>
<th>Link Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node: 2, <a href="http://dx.plos.org/10.1371/journal.pone.0115253">http://dx.plos.org/10.1371/journal.pone.0115253</a>; rel=&quot;pid&quot;</td>
<td><a href="http://dx.doi.org/10.1371/journal.pone.0115253">http://dx.doi.org/10.1371/journal.pone.0115253</a>; rel=&quot;pid&quot;</td>
</tr>
</tbody>
</table>
Signposting: Work in Progress

**Demo**

Input any HTTP URI of a scholarly article, and hit Get Headers to see its corresponding signposting headers.

![URI of PDF file](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253.pdf)

**Signposting Headers for the Landing Page Pattern**

<table>
<thead>
<tr>
<th>Resource</th>
<th>Link Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node: 5</td>
<td><a href="http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253.PDF">http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253.PDF</a></td>
</tr>
<tr>
<td>Node: 4</td>
<td><a href="http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253">http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253</a></td>
</tr>
<tr>
<td>Node: 6</td>
<td><a href="http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253.XML">http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253.XML</a></td>
</tr>
<tr>
<td>Node: 2</td>
<td><a href="http://dx.plos.org/10.1371/journal.pone.0115253">http://dx.plos.org/10.1371/journal.pone.0115253</a></td>
</tr>
</tbody>
</table>

Herbert Van de Sompel
COAR Annual Meeting, Vienna, Austria, 12/04/2016
Robust Links
Reference Rot

- Links to Web at Large resources are subject to Reference Rot:
  - Link Rot: Link stops working, e.g. HTTP 404 “Not Found”
  - Content Drift: Linked content changes over time
    - Possibly to the extent that it becomes no longer representative of the content that was initially referenced

http://hiberlink.org
Herbert van de Sompel

COAR Annual Meeting, Vienna, Austria, 12/04/2016

Martin Klein et al. (2014) Scholarly context not found. In: PLOS ONE
http://dx.doi.org/10.1371/journal.pone.0115253
Combating Reference Rot

① Create a snapshot of the referenced resource in one of many web archives that support on-demand archiving (manual, API):
   - archive.today
   - Internet Archive
   - perma.cc
   - webcitation.org

② Reference snapshots actionably by using:
   - Original URI
   - Snapshot URI
   - Date/Time of snapshot
   in order to maximize link robustness

http://robustlinks.mementoweb.org/about/
Reference Resources Actionably

- When referencing resources, use Link Decorations to convey Original URI, Snapshot URI, Date/Time

```
<a href="http://hiberlink.org"
   data-versionurl="https://archive.is/drFFu"
   data-versiondate="2015-11-16">
</a>

<a href="https://archive.is/drFFu"
   data-originalurl="http://hiberlink.org"
   data-versiondate="2015-11-16">
</a>

- Legitimate in HTML5
- Can be made actionable with JavaScript, e.g. robustlinks.js

Herbert Van de Sompel et al. (2015) Robust Links - Link Decorations
http://robustlinks.mementoweb.org/spec/
See Robust Links at Work

D-Lib Magazine

November/December 2015
Volume 21, Number 11/12
Table of Contents

Reminiscing About 15 Years of Interoperability Efforts

Herbert Van de Sompel
Los Alamos National Laboratory
herbertv@lanl.gov

Michael L. Nelson
Old Dominion University
mln@cs.odu.edu

DOI: 10.1045/november2015-vandesompel

https://dx.doi.org/10.1045/november2015-vandesompel
See Robust Links at Work

Robust Links

- Get near page creation date 2015-11-16
- Get near link date 2015-10-06
- Get from archive.is

With actionable attributes as per the Robust Links specification. "Robust Links" are created to be working towards time-stamped snapshots of the referenced resources. These robust links can be used to find an instance of original content based on the near creation date, an instance of the original content based on the near link date, or an instance of the original content that is available from an archive. The Robust Links specification leverages the Hiberlink project. It leverages the OAI-PMH (1999) protocol as a way to break ground for a universal adoption of e-print archives. The discoverability of e-prints, actually making them easier to discover, is an important goal of the OAI-PMH, a protocol for the recurrent exchange of metadata which was to an extent inspired by the Dienst protocol.
Conclusion

There is no real conclusion. There are insights:

• One doesn’t do interoperability because of interoperability but to enable cross-node applications that add value

• Establishing interoperability across a vast amount of nodes is a huge challenge. But meaningful levels of interoperability can be achieved via really basic approaches.

• Unfortunately, not even discovery is a solved problem (although the solution is available)
  • Anurag’s keynote is a real embarrassment for our community

Leading organizations and projects should promote web-centric cross-repository interoperability
Establishing New Levels of Interoperability for Web-Based Scholarship

Herbert Van de Sompel
Los Alamos National Laboratory
@hvdsomp

Cartoon by:
Patrick Hochstenbach