Compound Information Object Archive prototype

Luydmila Balakireva
Ryan Chute
Stephan Drescher
Herbert Van de Sompel [1]
Zhiwu Xie

Digital Library Research & Prototyping Team
Research Library
Los Alamos National Laboratory

[1] herbertv@lanl.gov
[1] http://public.lanl.gov/herbertv

This work was supported by the Andrew W. Mellon Foundation
and by NSF award number IIS-0430906 (Pathways)

Scenario

- Thanks to OAI-ORE, publishing named graphs to represent compound information objects has become mainstream.
- The Internet Archive collects named graphs (well, Resource Maps that is) as a way to preserve compound information objects.
- When collecting named graphs, the Internet Archive also collects representations of resources referenced in the named graphs.
- The Wayback Machine can now be used to get a glimpse of the state of a compound information object at some particular moment in time.

- In the prototype, named graphs from 2 authors are collected by the Internet Archive.

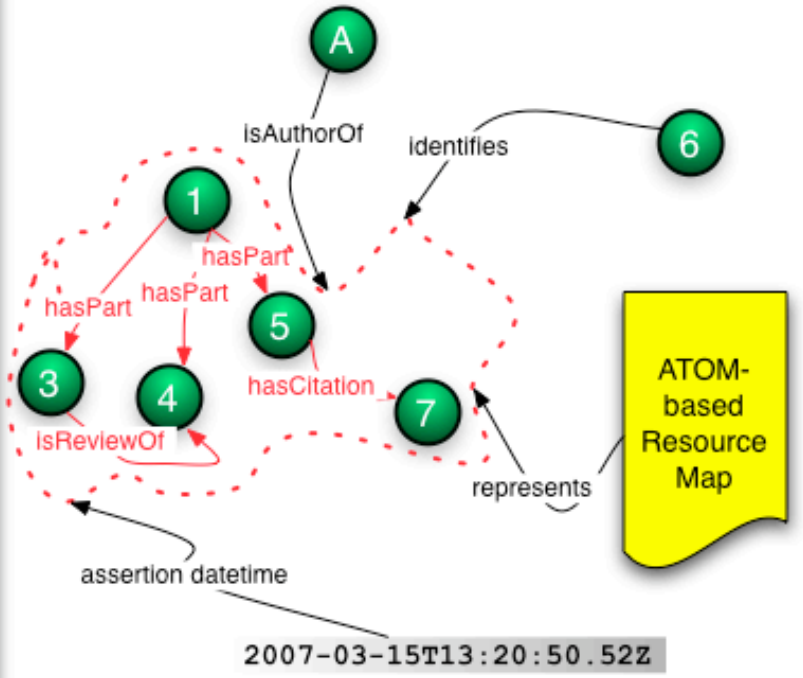
Issues Explored

- Named graphs
- Named graph serialization
- Named graph versioning
- Named graph containment node
- Compound object boundary
- Identification: named graph, containment node, work
- Referencing and re-use
- Named graph discovery

Technologies used in the prototype (1)

- ATOM feeds abused to serialize named graphs :-)
- Don't try this at home

```
<feed>

 <id>6</id>
  <link rel="self" href="6" type="application/ore+xml"/>
  <author>
    <uri>A</uri>
  </author>
  <updated>2007-03-15T13:20:50.52Z</updated>

 <entry>
  <id trix:uri="6">1</id>
  <link rel="hasPart" href="3" type="text/html"/>
  <link rel="hasPart" href="4" type="application/pdf"/>
  <link rel="hasPart" href="5" type="application/pdf"/>
 </entry>

 <entry>
  <id trix:uri="6">3</id>
  <link rel="isReviewOf" href="4" type="application/
                                                pdf"/>
 </entry>

 <entry>
  <id trix:uri="6">4</id>
 </entry>

 <entry>
  <id trix:uri="6">5</id>
  <link rel="hasCitation" href="7"/>
 </entry>

</feed>
```



(7) Not considered part of the compound object

Technologies used in the prototype (2)

- ATOM feeds abused to serialize named graphs :-)
  - HTTP URI that identifies the named graph => <id> element of ATOM <feed>
  - datetime of assertion of the named graph => <update> element of ATOM <feed>
  - author of the named graph => <author> element of ATOM <feed>

```
<?xml version="1.0" encoding="utf-8"?>
<feed xmlns="http://www.w3.org/2005/Atom" xmlns:trix="http://www.w3.org/2004/03/trix/trix-1/">
  <author>
    <uri>A</uri>
  </author>
  <updated>2007-03-15T13:20:50.52Z</updated>
  <id>6</id>
  <link rel="self" href="6" type="application/ore+xml" />
…
</feed>
```

Technologies used in the prototype (3)

- ATOM feeds abused to serialize named graphs :-)
  - Each quad of named graph is mapped to an <entry> of the feed:
    - Named graph identifier => trix:uri attribute to <entry>
    - URI of subject of triple => <id> of <entry>
    - URI of object of triple => @href of <link> element of <entry>
    - URI of predicate of triple => @rel of <link> element of <entry>

```
…
<entry>
  <id trix:uri="6">1</id>
  <link rel="self" href="1" type="text/html" />
  <link rel="info:lanl-repo/sem/ore/hasPart" href="3" type="text/html" />
  <link rel="info:lanl-repo/sem/ore/hasPart" href="4" type="application/pdf" />
  <link rel="info:lanl-repo/sem/ore/hasPart" href="5" type="application/pdf" />
…
</entry>
…
```

Technologies used in the prototype (4)

- ATOM feeds abused to serialize named graphs :-)
  - Named graph that re-uses (references) another named graph:
    - Use HTTP URI of the re-used named graph as <id> of the <entry>
    - A <link> element of <entry> points at the named graph serialization, recognizable by @type = application/ore+xml

```
…
<entry>
  <id trix:uri="6">Graph10</id>
  <link rel="alternate" type="application/ore+xml" href="Graph10" />
  <link rel="info:lanl-repo/sem/ore/isPartOf" href="5" />
…
</entry>
…
```

Technologies used in the prototype (5)

- ATOM feeds abused to serialize named graphs :-)
  - Resource that knows it is part of another named graph:
    - A <link> element of <entry> points at the other named graph:
      - HTTP URI of named graph => @href
      - isMemberOfGraph => @rel
      - application/ore+xml => @type

```
…
<entry>
  <id trix:uri="6">5</id>
  <link rel="info:lanl-repo/sem/ore/isMemberOfGraph" href="Graph20" type="application/ore+xml" />
  <link rel="info:lanl-repo/sem/ore/isPartOf" href="1"/>
…
</entry>
…
```

Technologies used in the prototype (6)

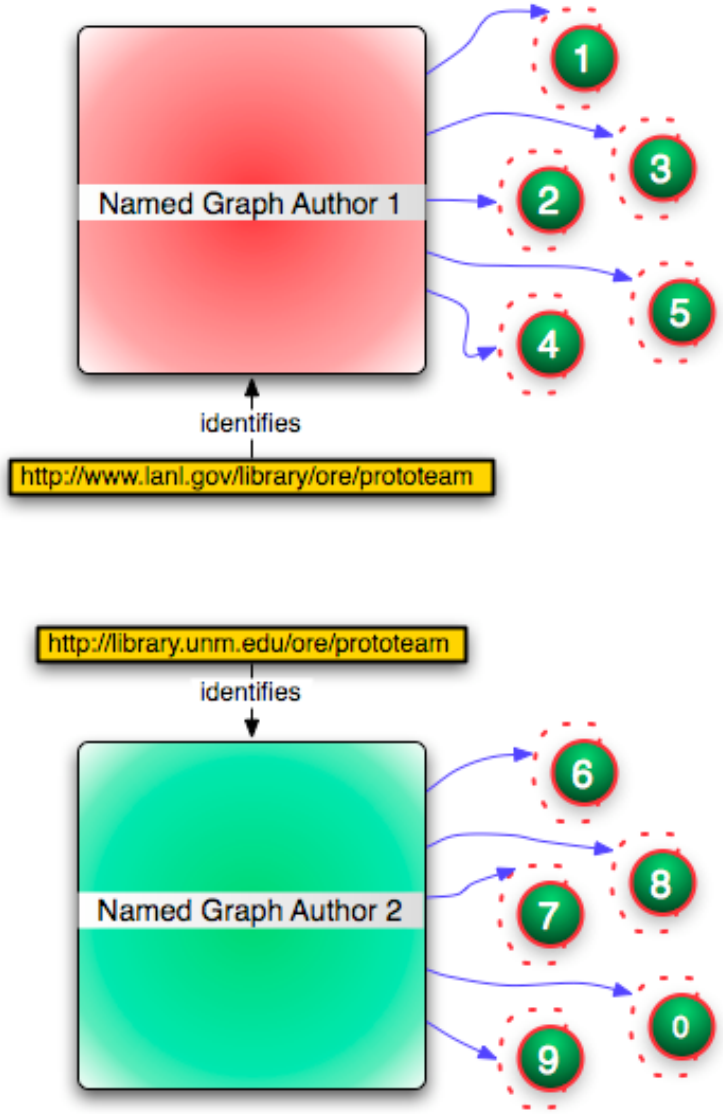
- Sitemaps to support named graph discovery
  - <loc> is HTTP URI of named graph
  - <lastmod> is datetime of assertion of named graph

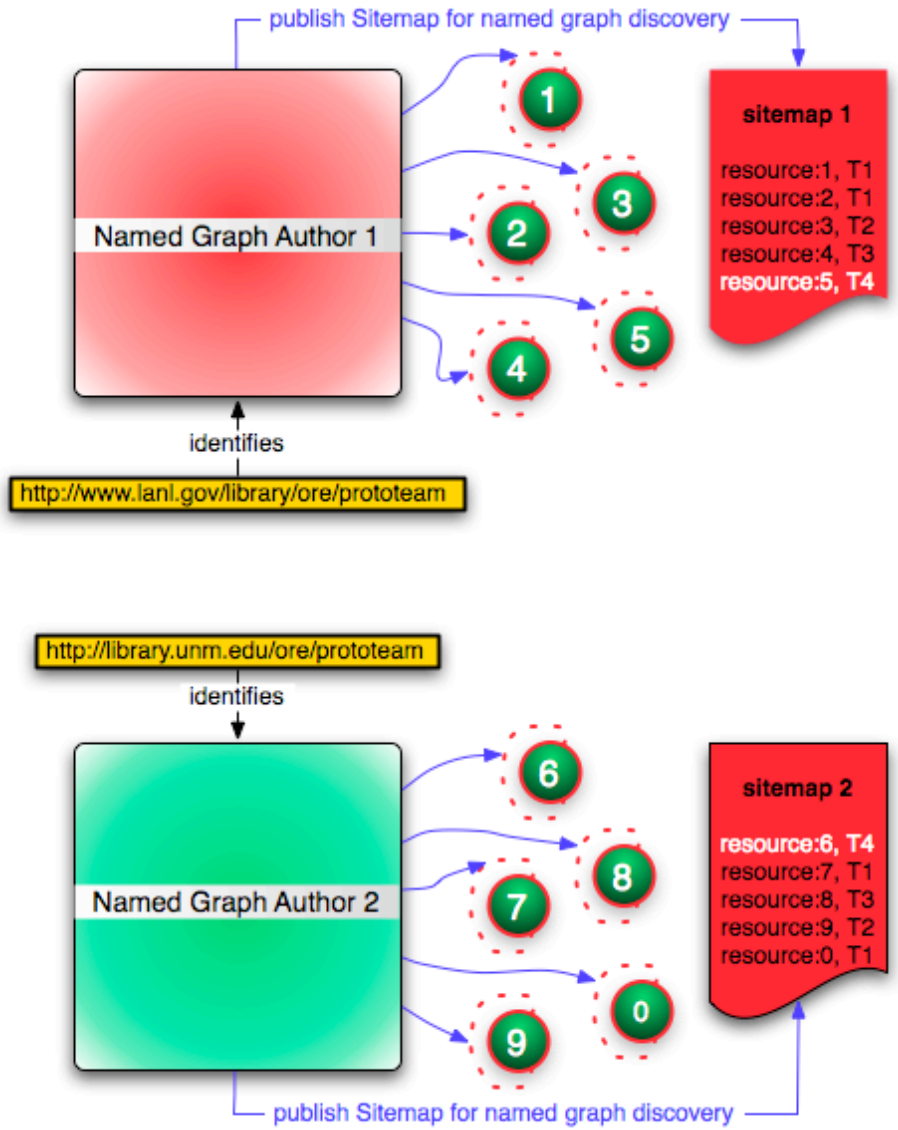
```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>6</loc>
    <lastmod> 2007-03-15T13:20:50.52Z</lastmod>
  </url>
  <url>
    <loc>Graph20</loc>
    <lastmod>2007-05-22T08:39:38Z</lastmod>
  </url>
…
</urlset>
```

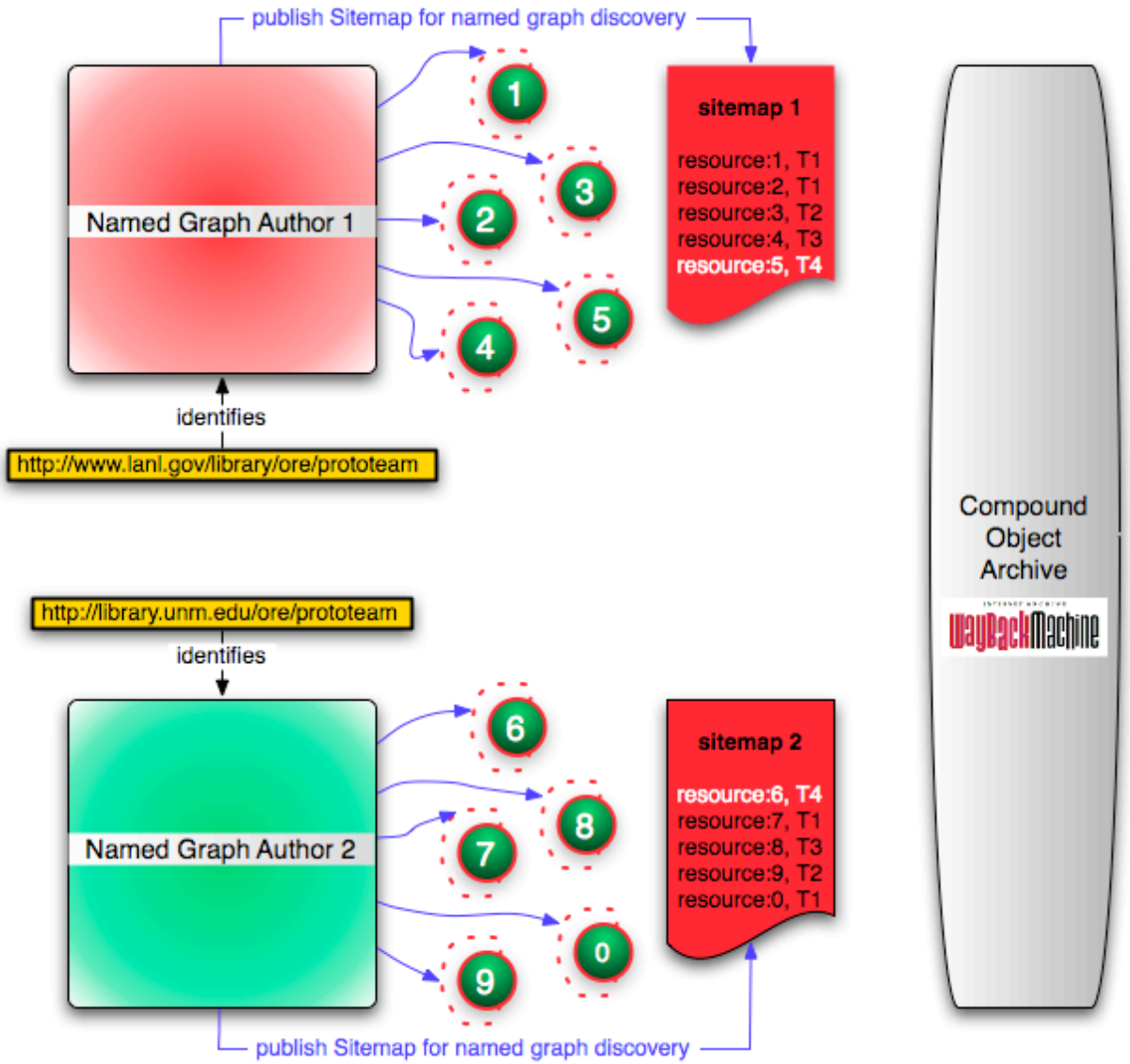
Technologies used in the prototype (7)

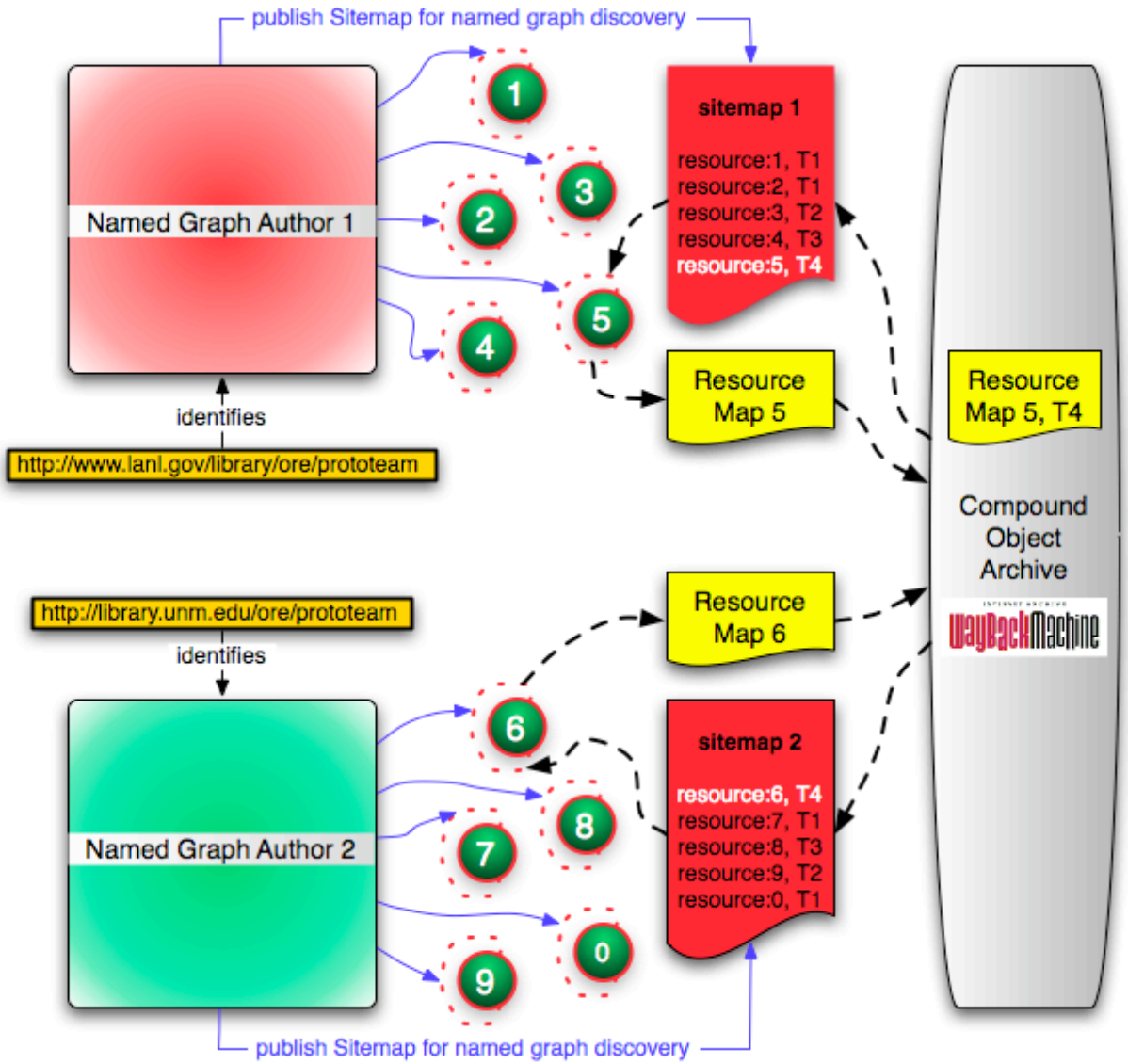
- Internet Archive Heritrix toolkit
- Simile Timeline
- Webdot graph visualization toolkit

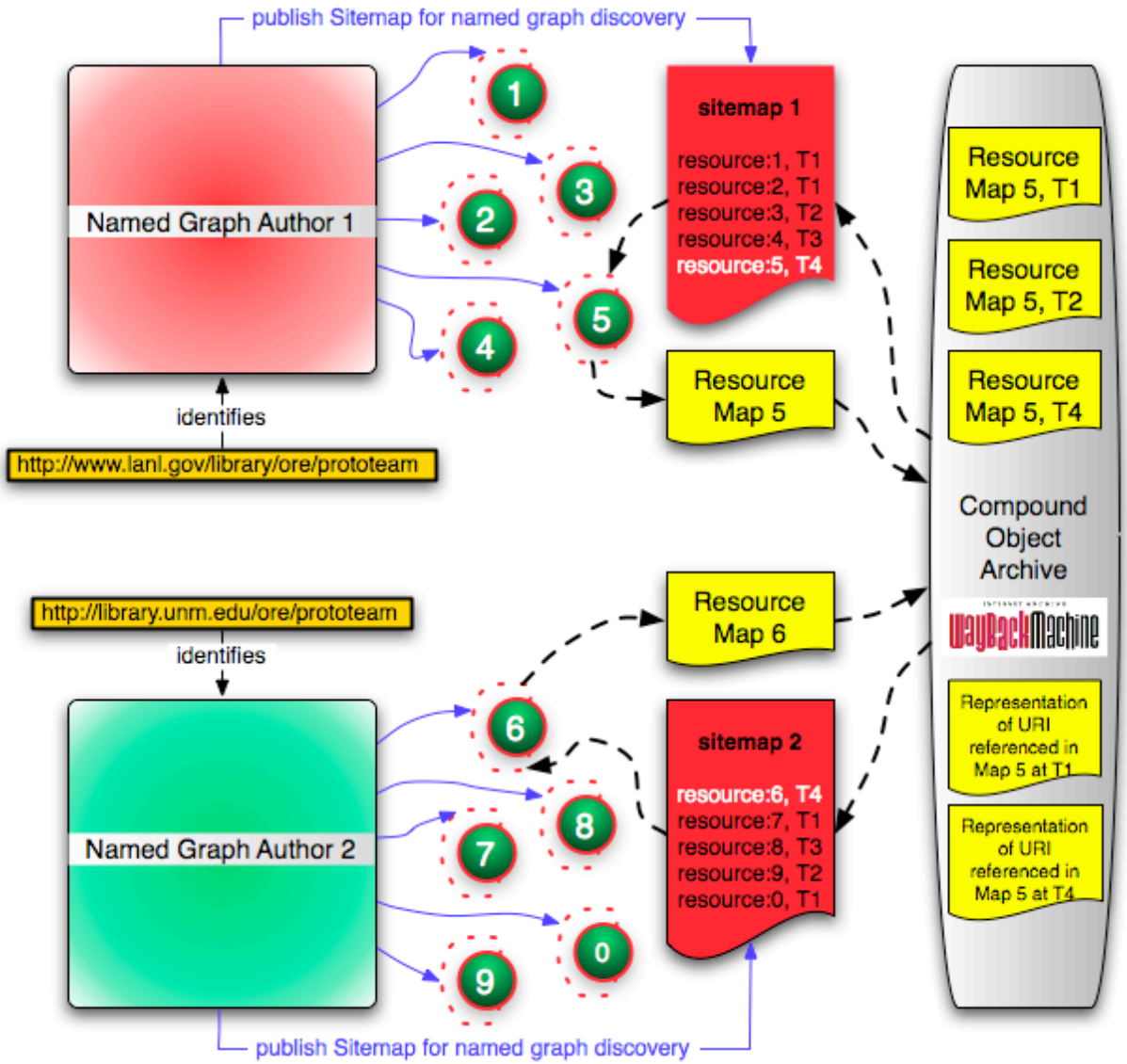
Prototype Environment

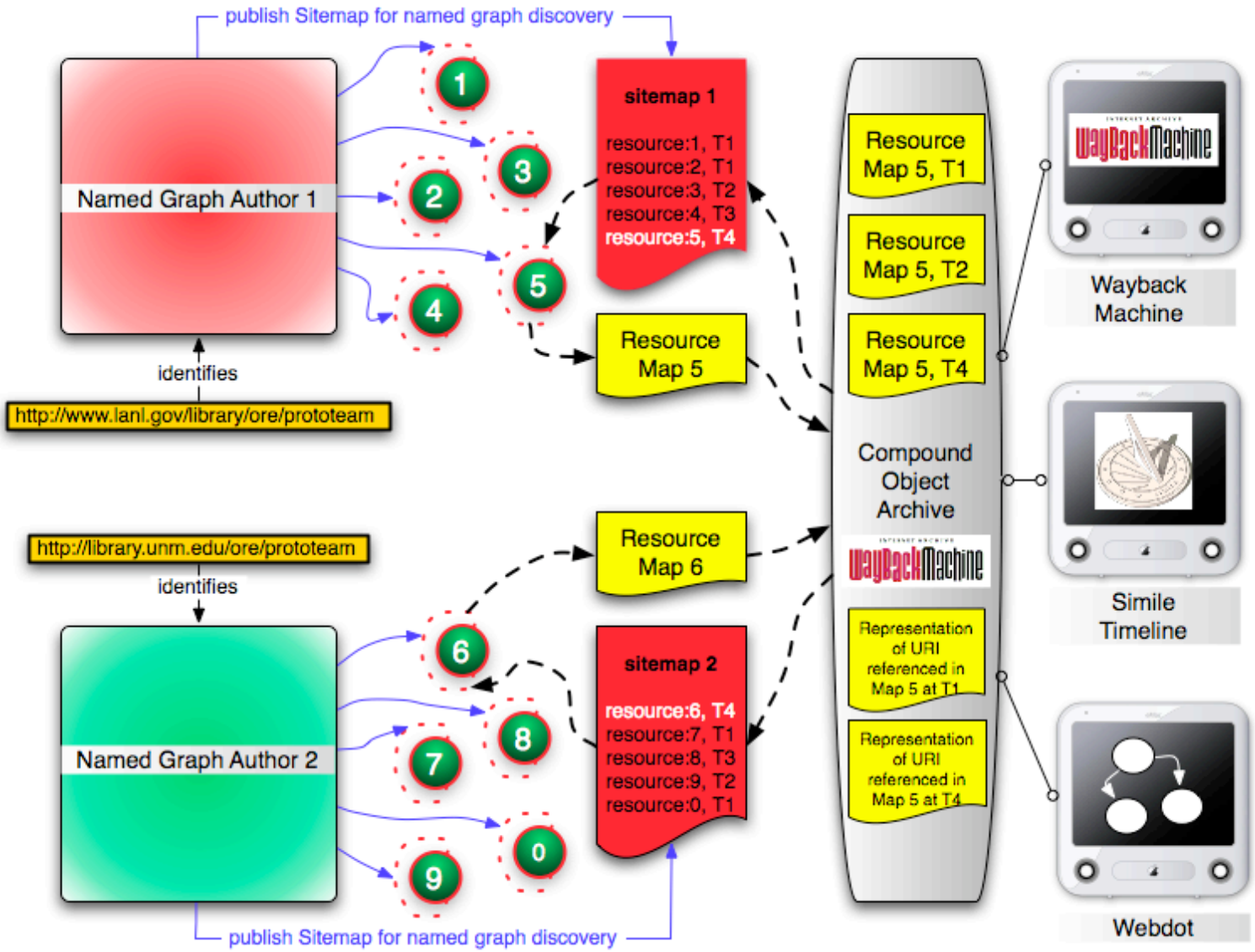

Named Graph Author 1
1
2
3
4
5
identifies
http://www.lanl.gov/library/ore/prototeam
http://library.unm.edu/ore/prototeam
identifies
Named Graph Author 2
6
7
8
9
0


publish Sitemap for named graph discovery
Named Graph Author 1
1
2
3
4
5
sitemap 1
resource:1, T1
resource:2, T1
resource:3, T2
resource:4, T3
resource:5, T4
identifies
http://www.lanl.gov/library/ore/prototeam
http://library.unm.edu/ore/prototeam
identifies
Named Graph Author 2
6
7
8
9
0
sitemap 2
resource:6, T4
resource:7, T1
resource:8, T3
resource:9, T2
resource:0, T1
publish Sitemap for named graph discovery

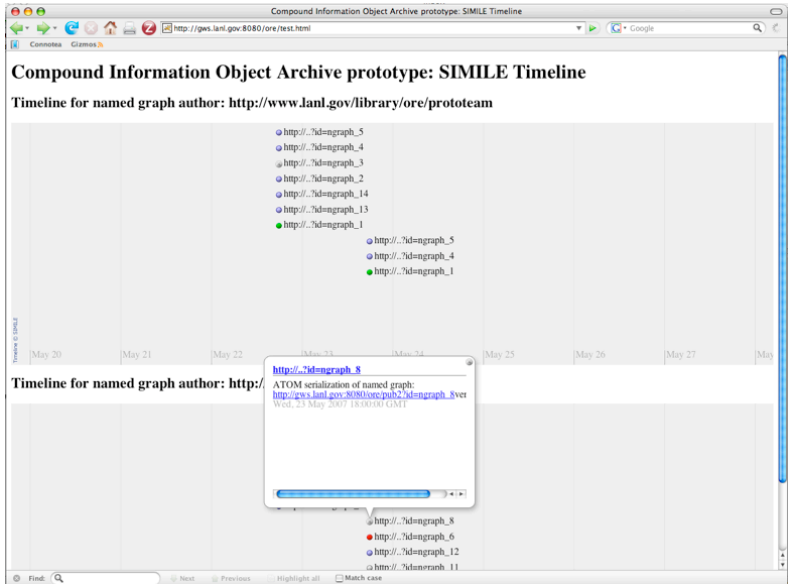

publish Sitemap for named graph discovery
Named Graph Author 1
1
2
3
4
5
sitemap 1
resource:1, T1
resource:2, T1
resource:3, T2
resource:4, T3
resource:5, T4
identifies
http://www.lanl.gov/library/ore/prototeam
Compound Object Archive
WayBackMachine
http://library.unm.edu/ore/prototeam
identifies
Named Graph Author 2
6
7
8
9
0
sitemap 2
resource:6, T4
resource:7, T1
resource:8, T3
resource:9, T2
resource:0, T1
publish Sitemap for named graph discovery

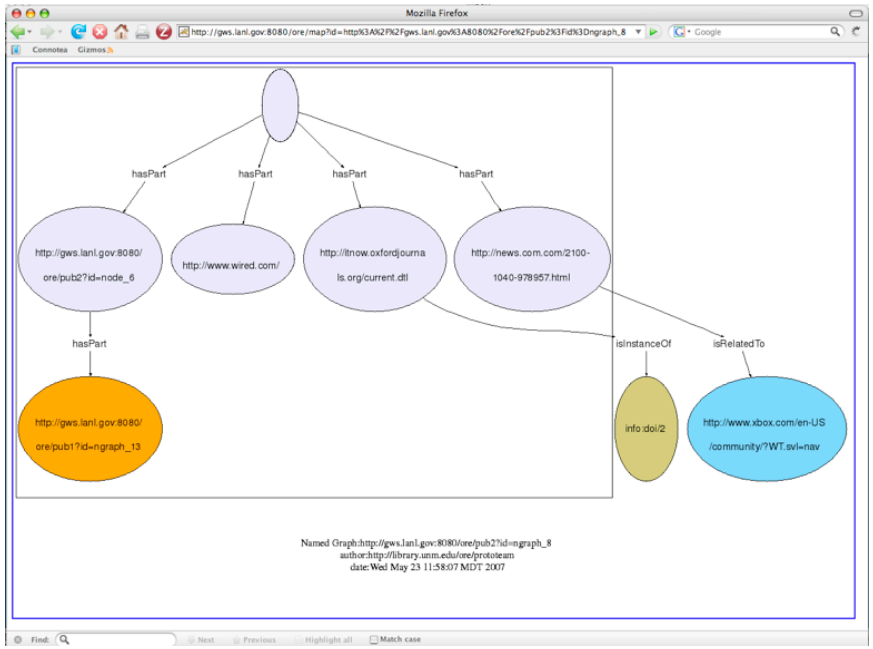

publish Sitemap for named graph discovery
Named Graph Author 1
identifies
http://www.lanl.gov/library/ore/prototeam
1
2
3
4
5
sitemap 1
resource:1, T1
resource:2, T1
resource:3, T2
resource:4, T3
resource:5, T4
Resource Map 5
Resource Map 5, T4
Compound Object Archive
WayBackMachine
http://library.unm.edu/ore/prototeam
identifies
Named Graph Author 2
6
7
8
9
0
Resource Map 6
sitemap 2
resource:6, T4
resource:7, T1
resource:8, T3
resource:9, T2
resource:0, T1
publish Sitemap for named graph discovery

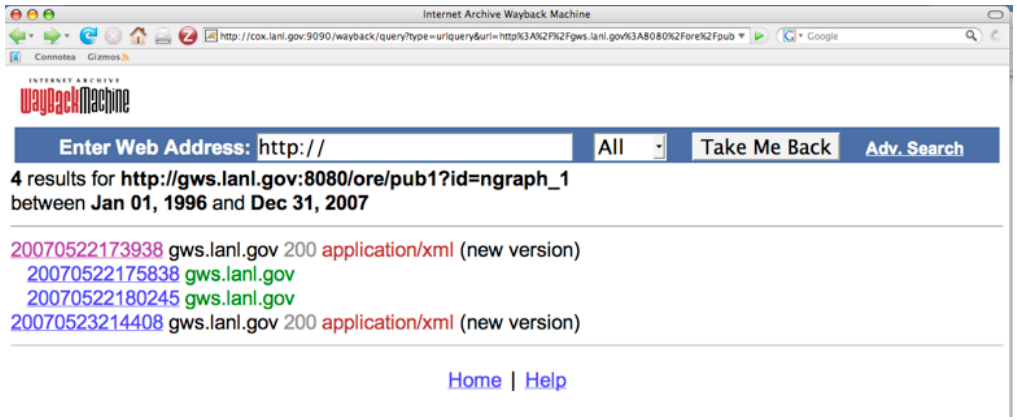

publish Sitemap for named graph discovery
Named Graph Author 1
identifies
http://www.lanl.gov/library/ore/prototeam
1
2
3
4
5
sitemap 1
resource:1, T1
resource:2, T1
resource:3, T2
resource:4, T3
resource:5, T4
Resource Map 5
Resource Map 5, T1
Resource Map 5, T2
Resource Map 5, T4
Compound Object Archive
WayBackMachine
Representation of URI referenced in Map 5 at T1
Representation of URI referenced in Map 5 at T4
http://library.unm.edu/ore/prototeam
identifies
Named Graph Author 2
6
7
8
9
0
Resource Map 6
sitemap 2
resource:6, T4
resource:7, T1
resource:8, T3
resource:9, T2
resource:0, T1
publish Sitemap for named graph discovery


publish Sitemap for named graph discovery
Named Graph Author 1
identifies
http://www.lanl.gov/library/ore/prototeam
1
2
3
4
5
sitemap 1
resource:1, T1
resource:2, T1
resource:3, T2
resource:4, T3
resource:5, T4
Resource Map 5
http://library.unm.edu/ore/prototeam
identifies
Named Graph Author 2
6
7
8
9
0
Resource Map 6
sitemap 2
resource:6, T4
resource:7, T1
resource:8, T3
resource:9, T2
resource:0, T1
publish Sitemap for named graph discovery
Resource Map 5, T1
Resource Map 5, T2
Resource Map 5, T4
Compound Object Archive
Representation of URI referenced in Map 5 at T1
Representation of URI referenced in Map 5 at T4
Wayback Machine
Simile Timeline
Webdot

Movie of Prototype







Los Alamos NATIONAL LABORATORY

OPEN ARCHIVES

Findings regarding the explored issues (1)

- Named graphs
  - A truly powerful concept. It just feels right.
- Named graph serialization
  - ATOM was fun to play with, works to an extent, but really using it for named graph representation is probably a far stretch.
  - Can not do literals
  - Has <extension> mechanism
  - I trust the semantic web community would think we're total idiots
- Named graph versioning
  - Datetime of named graph assertion is handy for consuming applications
  - Question remains exactly under which conditions the datetime changes

Findings regarding the explored issues (2)

- Named graph containment node
  - Starting to think to always use a blank node:
    - No ambiguity in referencing: the HTTP URI of the named graph is the reference.
    - What would be a realistic candidate for this node anyhow? An existing resource? Which one? The splash page? Isn't the splash page just another Part of the CO?

- Identification: named graph, containment node, work
  - HTTP URI is great for named graph ~ discovery of Resource Map
  - Containment node: blank?
  - *Work* identifier: provide as isIntanceOf

Findings regarding the explored issues (3)

- Referencing and re-use:
  - HTTP URI of Named Graph is da handle!
  - If one would use the URI of the containment node, then how does that express the CO? OK, via some discovery mechanism. But there could be multiple named graphs (corresponding with the same CO) that share the containment node. OK, so lets create unique containment nodes per named graph. So, now we have 2 unique identifiers per named graph: HTTP URI of named graph & URI of containment node. Too much to be good?
  - And what would it mean for the URI of the containment node to be the URI of a named graph, i.e. containment node is a named graph. Bizar, right!? Meaning we should now state that URI of containment node can not be URI of a named graph. Hum …

- Named graph discovery:
  - Anything goes that supports discovery.
  - Promote a few approaches.
  - I feel OAI-PMH has something to offer, especially in light of named graph versioning and installed base.