Open Archives Initiative
Object Reuse & Exchange

Resource Map Discovery

Michael L. Nelson*
Carl Lagoze, Herbert Van de Sompel, Pete Johnston, Robert Sanderson, Simeon Warner

OAI-ORE Specification Roll-Out
Open Repositories 2008
Southampton UK, April 4, 2008

*Old Dominion University, Norfolk VA
http://www.cs.odu.edu/~mln/

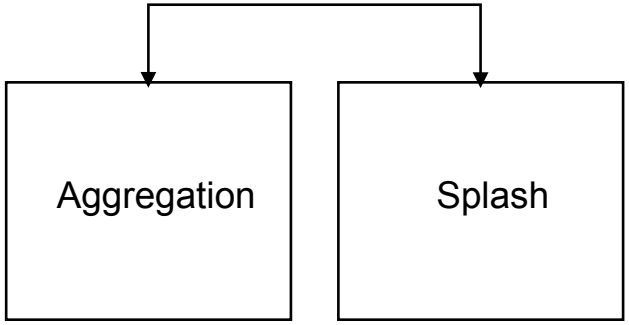
Discovery…

Resource Map Discovery Outline

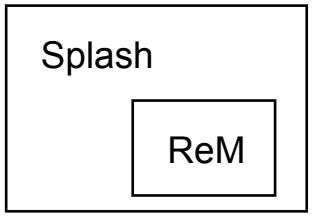
- Different Idioms for Design & Discovery
- Batch Discovery
  - OAI-PMH, SiteMaps, RSS/Atom
- Embedding Discovery Links
  - With HTML “link” element
  - With HTTP “Link” response header
- Open Issues
  - Indirect HTML/HTTP discovery links
  - Proxies from the ADM
    - HTML support for using “Proxy” (aka “cite in context”)
    - The nature of URI-P (resolvable vs. non-resolvable)

Different Design Methods, Different Discovery Methods

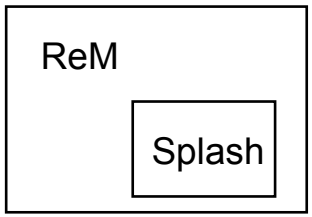
Syndication Format Idiom
(URI-A ≠ URI-S)



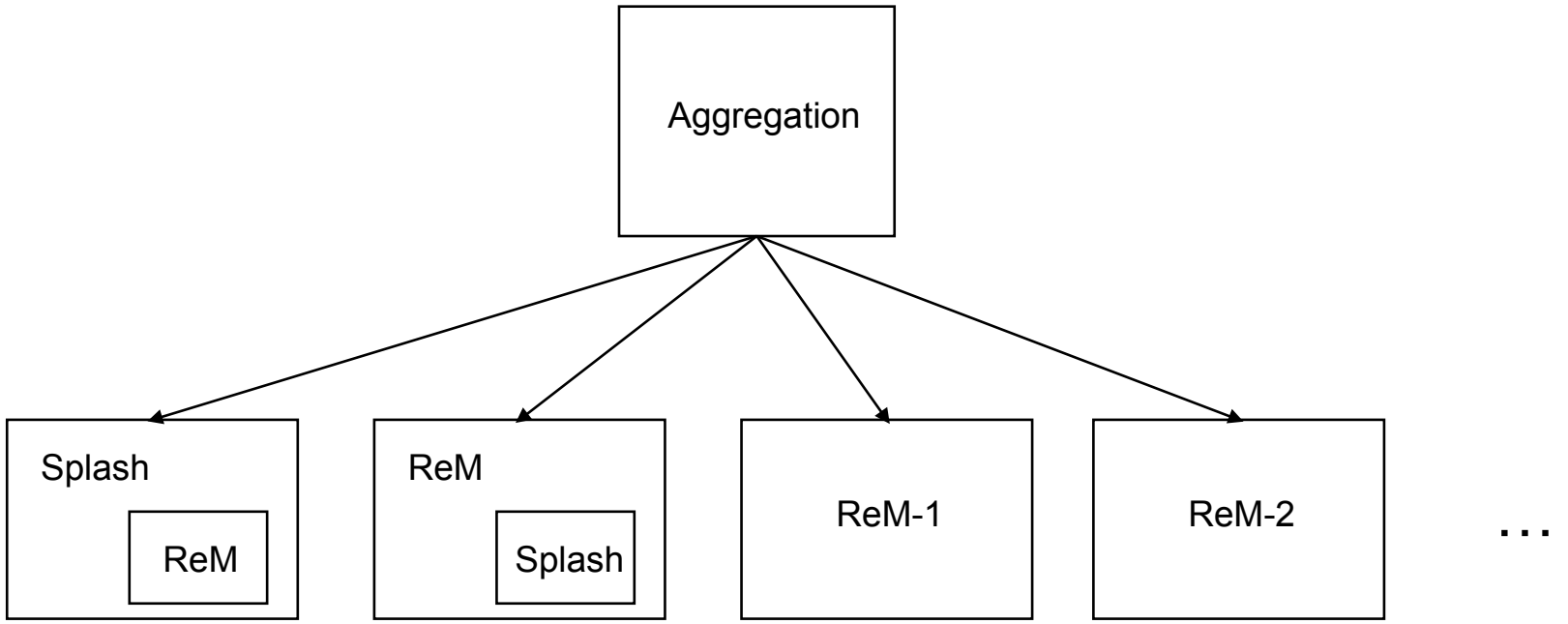
RDFa / Microformat Idiom
(URI-R = URI-S)



XML Stylesheet Idiom
(URI-R = URI-S)



Putting it All Together…



Multiple values for URI-R and URI-S.
The only unique value is URI-A.

Batch Discovery

- ReMs & Aggregations are resources and we already know how to expose large batches of resources:
  - OAI-PMH
  - SiteMaps
  - RSS/Atom

Batch :: OAI-PMH

http://www.foo.edu/oai?verb=ListRecords&metadataPrefix=oai_rem_atom

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
         http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2007-02-08T08:55:46Z</responseDate>
  <request verb=”ListRecords” metadataPrefix="oai_rem_atom">
        http://foo.edu/oai2</request>
  <ListRecords>
   <record>
    <header>
      <identifier>oai:foo.edu:object1</identifier>
      <datestamp>2007-01-06</datestamp>
    </header>
    <metadata>
        <!-- Insert object1 ReM here -->
    </metadata>
  </record>
  . . .
 </ListRecords>
</OAI-PMH>
```

MUST NOT
equal either ReM Atom /feed/id or
/feed/link[@rel="self"]/@href

MUST be equal to ReM Atom /feed/updated

OAI-PMH GetRecord Processing

http://www.foo.edu/oai?verb=GetRecord&identifier=oai:foo.edu:object1&metadataPrefix=oai_rem_atom

http://some.gateway.org/pmh2ore?=http://foo.edu/oai2?verb=GetRecord&metadataPefix=oai_rem_atom&identifier=oai:foo.edu:object1

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
         http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2007-02-08T08:55:46Z</responseDate>
  <request verb="GetRecord" identifier="oai:foo.edu:object1"
           metadataPrefix="oai_rem_atom">http://foo.edu/oai2</request>
  <GetRecord>
   <record>
    <header>
      <identifier>oai:foo.edu:object1</identifier>
      <datestamp>2007-01-06</datestamp>
    </header>
    <metadata>
        <!-- Insert Object1 ReM here -->
    </metadata>
  </record>
 </GetRecord>
</OAI-PMH>
```

need a gateway to:
1. strip off OAI-PMH wrappers
2. return just what is inside <metadata>
3. reset the MIME type (e.g., from application/xml to application/atom+xml)

Batch :: SiteMaps

http://www.foo.edu/sitemap-rem.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.foo.edu/objects/object1.atom</loc>
      <lastmod>2007-01-06</lastmod>
   </url>
   <url>
      <loc>http://www.foo.edu/objects/object2.atom</loc>
      <lastmod>2007-08-11</lastmod>
      <changefreq>weekly</changefreq>
   </url>
   <url>
      <loc>http://www.foo.edu/objects/object3.atom</loc>
      <lastmod>2007-03-15T18:30:02Z</lastmod>
      <priority>0.3</priority>
   </url>
...
</urlset>
```

MUST equal /feed/link[@rel="self"]/@href or /feed/id for corresponding ReM

MUST be equal to ReM Atom /feed/updated

remember SiteMap path limitation: http://www.foo.edu/a/b/sitemap-rem.xml can list http://www.foo.edu/a/b/bar2.atom but not http://www.foo.edu/bar1.atom

Batch :: RSS

http://www.foo.edu/all-rems.rss

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>ReMs at www.foo.edu</title>
    <link>http://www.foo.edu/</link>
    <description>All of the Resource Maps for resources at www.foo.edu</description>

    <item>
      <title>ReM for Object 1</title>
      <link>http://www.foo.org/objects/object1.atom</link>
      <description>ReM for Object 1</description>
      <pubDate>Sat, 06 Jan 2007 00:00:00 GMT</pubDate>
    </item>

    <item>
      <title>ReM for Object 2</title>
      <link>http://www.foo.org/objects/object2.atom</link>
      <description>ReM for Object 2</description>
      <pubDate>Sat, 11 Aug 2007 00:00:00 GMT</pubDate>
    </item>
</channel>
</rss>
```

MUST equal ReM Atom /feed/id;
or /feed/link[@rel="self"]/@href

MUST equal ReM Atom /feed/updated
(after conversion from RFC-822 format to ISO 8601 format)

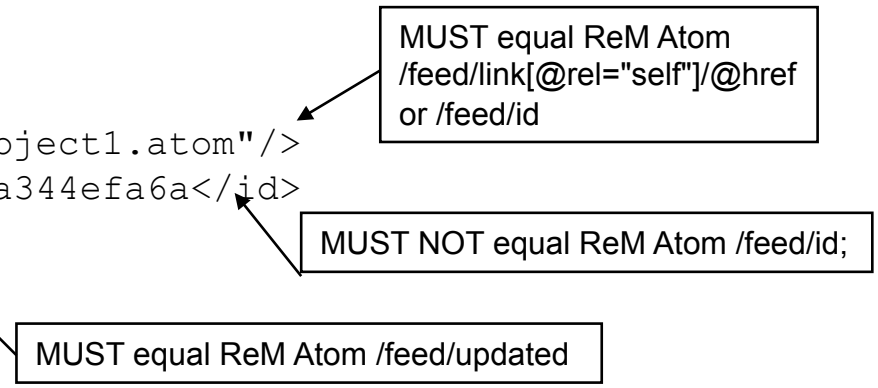
Batch :: Atom

http://www.foo.edu/all-rems.atom

```
<feed xmlns="http://www.w3.org/2005/Atom">
 <title>ReMs at www.foo.edu</title>
 <link href="http://www.foo.edu/" />
 <link href="http://www.foo.edu/all-rems.atom" rel="self"/>
 <updated>2007-08-15T18:30:02Z</updated>
 <author>
   <name>John Doe</name>
   <email>johndoe@foo.edu</email>
 </author>
 <id>urn:uuid:60a76c80-d399-11d9-b91C-0003939e0af6</id>

 <entry>
   <title>ReM For Object1</title>
   <link href="http://www.foo.org/objects/object1.atom"/>
   <id>urn:uuid:1225c695-cfb8-4ebb-aaaa-80da344efa6a</id>
   <updated>2007-01-06T00:00:00Z</updated>
 </entry>

 <entry>
   <title>ReM For Object2</title>
   <link href="http://www.foo.org/objects/object2.atom"/>
   <id>urn:uuid:9a2cc699-ccba-9e8b-132e-91da394e9a5c</id>
   <updated>2007-08-11T00:00:00Z</updated>
 </entry>
</feed>
```



Embedding Discovery Links into Resources

- Starting with a resource, how to find the associated Aggregations(s)?
  - HTML `<link>`
  - HTTP Response Headers
  - HTTP `<A>` & `<IMG>`
  - ReM Transparency
- 4 levels to describe resources' knowledge of their Aggregations

Embedding :: Knowledge Levels

- Full knowledge
  - the Aggregation(s) is linked to by all resources in the Aggregation.
- Indirect knowledge
  - all but one of the resources in the Aggregation link to a single, unique Aggregated Resource in the aggregation, which in turn links to the Aggregation(s).
  - functionally the same as full knowledge, but likely to be useful in actual deployment
- Limited knowledge
  - only a subset of the resources in the Aggregation (typically just a single resource) link to the Aggregation(s), and the remainder of the resources have no links at all.
- Zero knowledge
  - none of the resources in a Aggregation link to the Aggregation.

HTML <link> :: Full Knowledge

```
<html>
<head>
<title>Hello World.</title>
<link href="http://example.net/hw.atom#aggregation"
      type="application/atom+xml" rel="resourcemap" >
</head>
<body>
<img src="hello.jpeg">
<img src="world.jpeg">
</html>
```

HTML <link> :: Indirect Knowledge

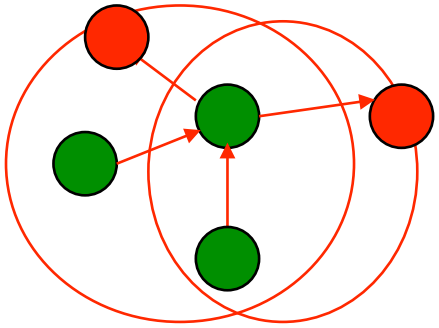
```
<html>
<head>
<title>Chapter Twelve.</title>
<link href="http://mybook.com/toc.html"
      type="text/html" rel="indirectresourcemap" >
</head>
<body>
Welcome to chapter twelve...
</body>
</html>
```

Open Issue: In certain scenarios, indirect linking could be useful.
In other scenarios, it can lead to incorrect assertions:

T_1: A-1 aggregates {AR-1, AR-2, AR-3}.
AR-1 directly links to A-1.
AR-2, AR-3 indirectly link through AR-1.

T_2: A-2 aggregates {AR-1, AR-2}.

T_3: AR-1 updates its direct links to include {A-1, A-2}
AR-2 is telling the truth, but AR-3 is not.



Embedding :: HTTP Response

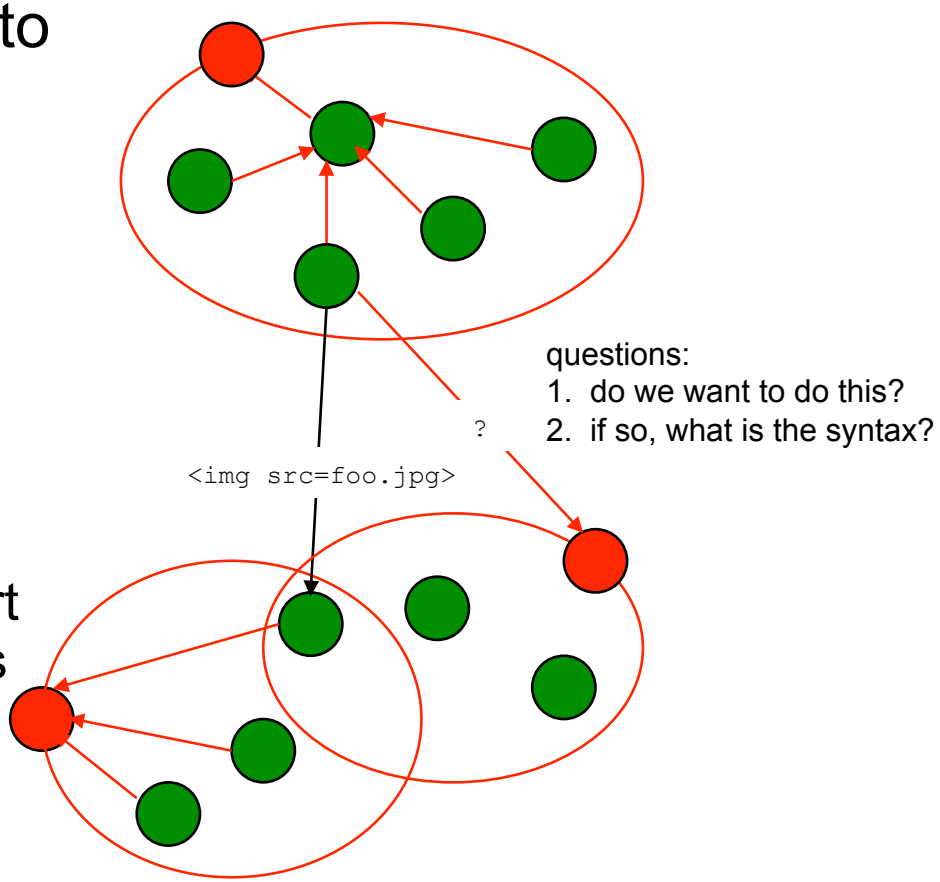
```
HEAD http://www.example.net/hello.jpeg HTTP/1.1
Host: www.example.net
Connection: close


HTTP/1.1 200 OK
Date: Sat, 26 May 2007 22:43:10 GMT
Server: Apache/2.2.0
Last-Modified: Sat, 26 May 2007 19:32:04 GMT
ETag: "c3596-816-92123500"
Accept-Ranges: bytes
Content-Length: 2070
Link: <http://example.net/hw.atom>; type="application/atom+xml"; rel="resourcemap"
Content-Type: image/jpeg
Connection: close
```

Nottingham's IETF Draft establishing semantic equivalence between HTML `<link>` and HTTP `Link:`

HTML `<link>` vs. `<A>` & `<IMG>`

- `link` is from “this” document to its 1 or more corresponding ReMs
  - AR-1 ore:isAggregatedBy A-1
- `A` & `IMG` capabilities are proposed to provide “hints” about the context of the disaggregated resources
  - problem: HTML does not support statements of the form “I got this from there”
  - example: “I got this JPEG from ReM1, the PDF from ReM2 and this quoted text section from ReM3.”



HTML Option #1: `resourcemap` attribute

```
<html>
...
Here is a helpful reference for distinguishing
<a href="http://example.org/pics/f-t.pdf"
resourcemap="http://example.org/amphibians.atom">frogs vs. toads</a>.
<p>
Here is a frog
<img src="http://weluvfrogs.org/imgs/frog12.jpeg"
resourcemap="http://frogs.org/frogs.atom">
and here is a toad <img src="http://toadsrule.org/toad.gif"
resourcemap="http://toadsrule.org/toads.atom">.
...
</html>
```

Pro: very simple, human readable
Con: invalid HTML

HTML Option #2: <A> rel attribute

```
<html>
...
Here is a helpful reference for distinquishing
<a href="http://example.org/pics/f-t.pdf"
rel="resourcemap=http://example.org/amphibians.atom">frogs vs. toads</a>.
<p>
Here is a frog
<a rel="resourcemap=http://frogs.org/frogs.atom">
<img src="http://weluvfrogs.org/imgs/frog12.jpeg">
</a> and here is a toad
<a rel="resourcemap=http://toadsrule.org/toads.atom">
<img src="http://toadsrule.org/toad.gif">
</a>.
...
</html>
```

Pro: Valid HTML

Con: Not uniform (<A> and <IMG> do not (yet) support the same elements)

HTML Option #3: `class` attribute

```
<html>
...
Here is a helpful reference for distinguishing
<a href="http://example.org/pics/f-t.pdf"
class="resourcemap=http://example.org/amphibians.atom">frogs vs. toads</a>.
<p>
Here is a frog
<img src="http://weluvfrogs.org/imgs/frog12.jpeg"
class="resourcemap=http://frogs.org/frogs.atom">
and here is a toad <img src="http://toadsrule.org/toad.gif"
class="resourcemap=http://toadsrule.org/toads.atom">.
...
</html>
```

Pro: very simple, human readable, valid HTML

Con: stretches, but does not break, "`class`"*

* http://www.w3.org/TR/REC-html40/struct/global.html#adef-class

The class attribute has several roles in HTML:

* As a style sheet selector (when an author wishes to assign style information to a set of elements).
* For general purpose processing by user agents.

HTML Option #4: / elements

```
<html>
...
Here is a helpful reference for distinguishing
<span class="resourcemap=http://example.org/amphibians.atom">
<a href="http://example.org/pics/f-t.pdf" frogs vs. toads</a>.
</span>
<p>
Here is a frog
<span class="resourcemap=http://frogs.org/frogs.atom">
<img src="http://weluvfrogs.org/imgs/frog12.jpeg">
</span> and here is a toad
<span class="resourcemap=http://toadsrule.org/toads.atom">
<img src="http://toadsrule.org/toad.gif">
</span>.
...
</html>
```

Pro: Valid HTML, Uniform Approach, Consistent with Microformats (COinS, unAPI, etc.), can cite blocks of text (URIs resolved prior to HTML generation)
Con: The beginnings of a ReM Microformat… allow incomplete ReMs?

HTML Option #5: *Really* use URI-P

```
<html>
...
Here is a helpful reference for distinguishing
<!-- Option 1: concatenate (URI-A,URI-AR); not necessarily a "real" registry (redirect could be
scripted).  Shown w/o necessary encoding. -->
<a
href="http://purl.org/ore/http://example.org/amphibians.atom/*/http://example.org/pics/f-t.pdf">
frogs vs. toads</a>.
<p>
Here is a frog
<!-- Option 2: IA-like approach.  Number is a persistent offset into a "real" registry
(e.g., 12th Aggregation to aggregate this AR). 303 Redirect to URI-AR, put URI-A
in an ore:isAggregatedBy HTTP response header. No encoding needed. -->
<img src="http://purl.org/ore/12/http://weluvfrogs.org/imgs/frog12.jpeg">
and here is a toad
<!-- Option 3: TinyURL-like approach. Similar to above, but w/ compressed, semantic-free URI.  -->
<img src="http://purl.org/ore/2kjl87j">.
...
</html>
```

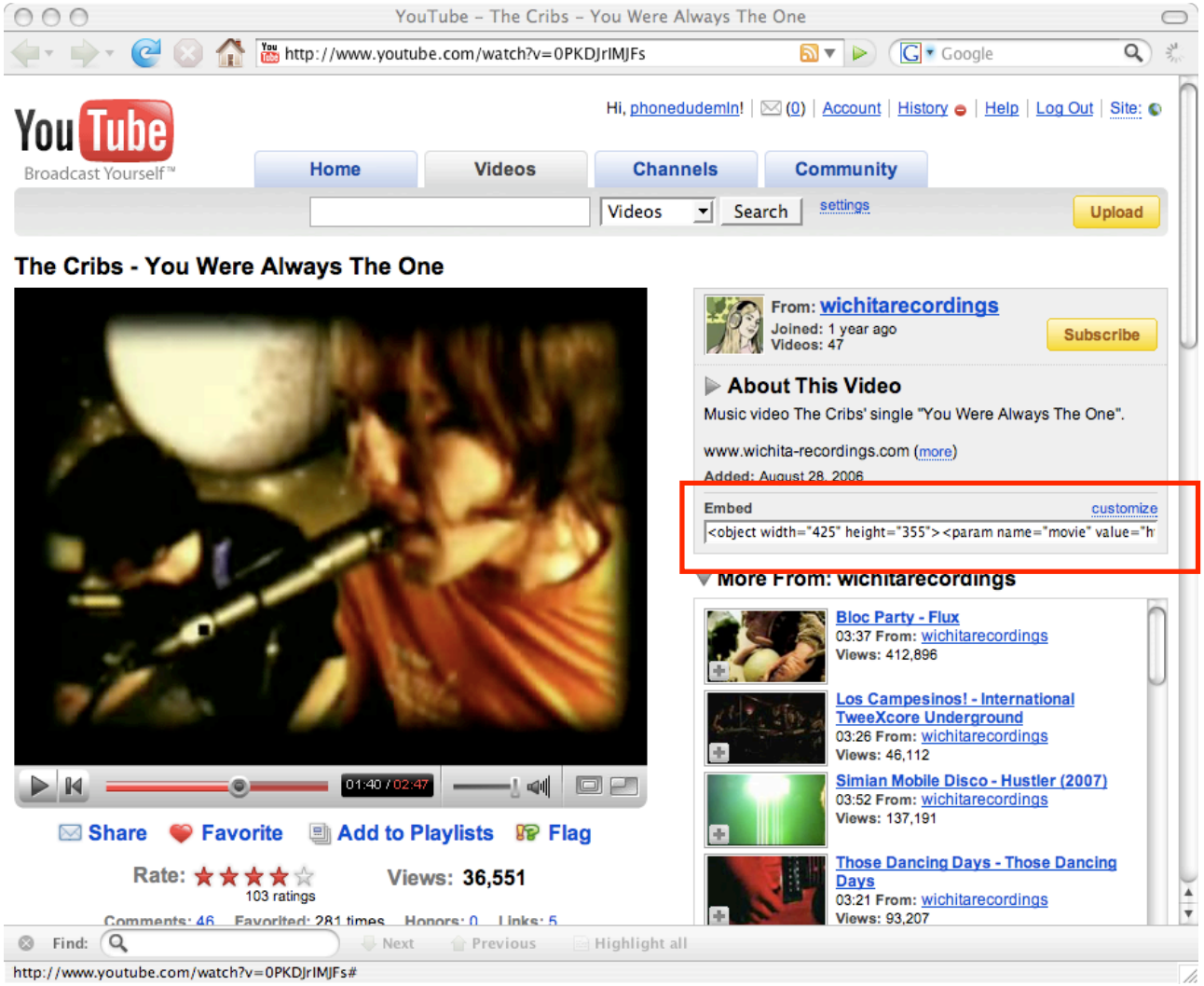
Pro: No HTML tricks.

Con: URI tricks, "some assembly required".

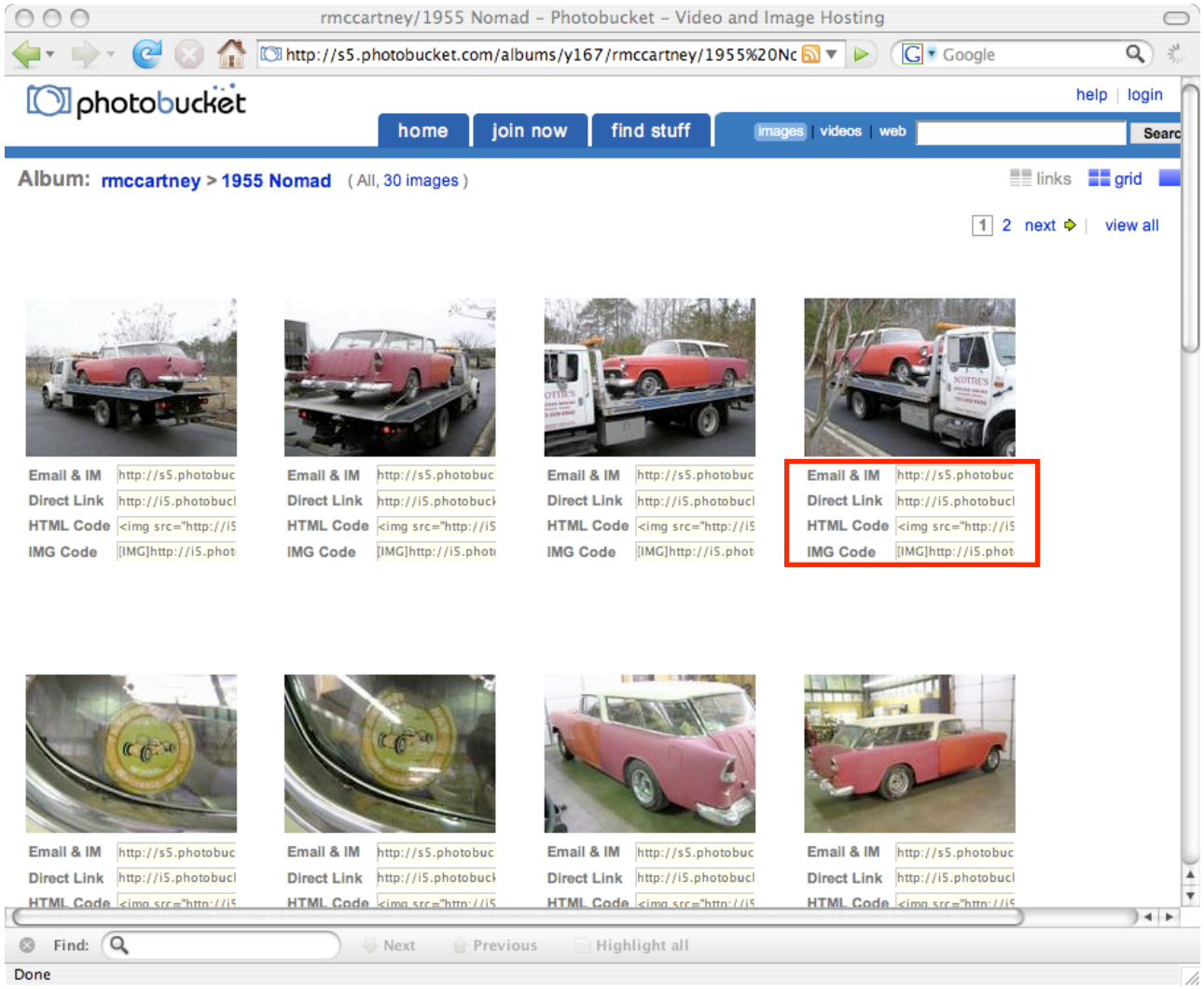
Embedding :: Transparency

- There is precedent for exposing URIs, HTML, JavaScript, etc. as opaque strings for users to paste into other applications
- This is not the same as creating a hypertext link to the scripts…

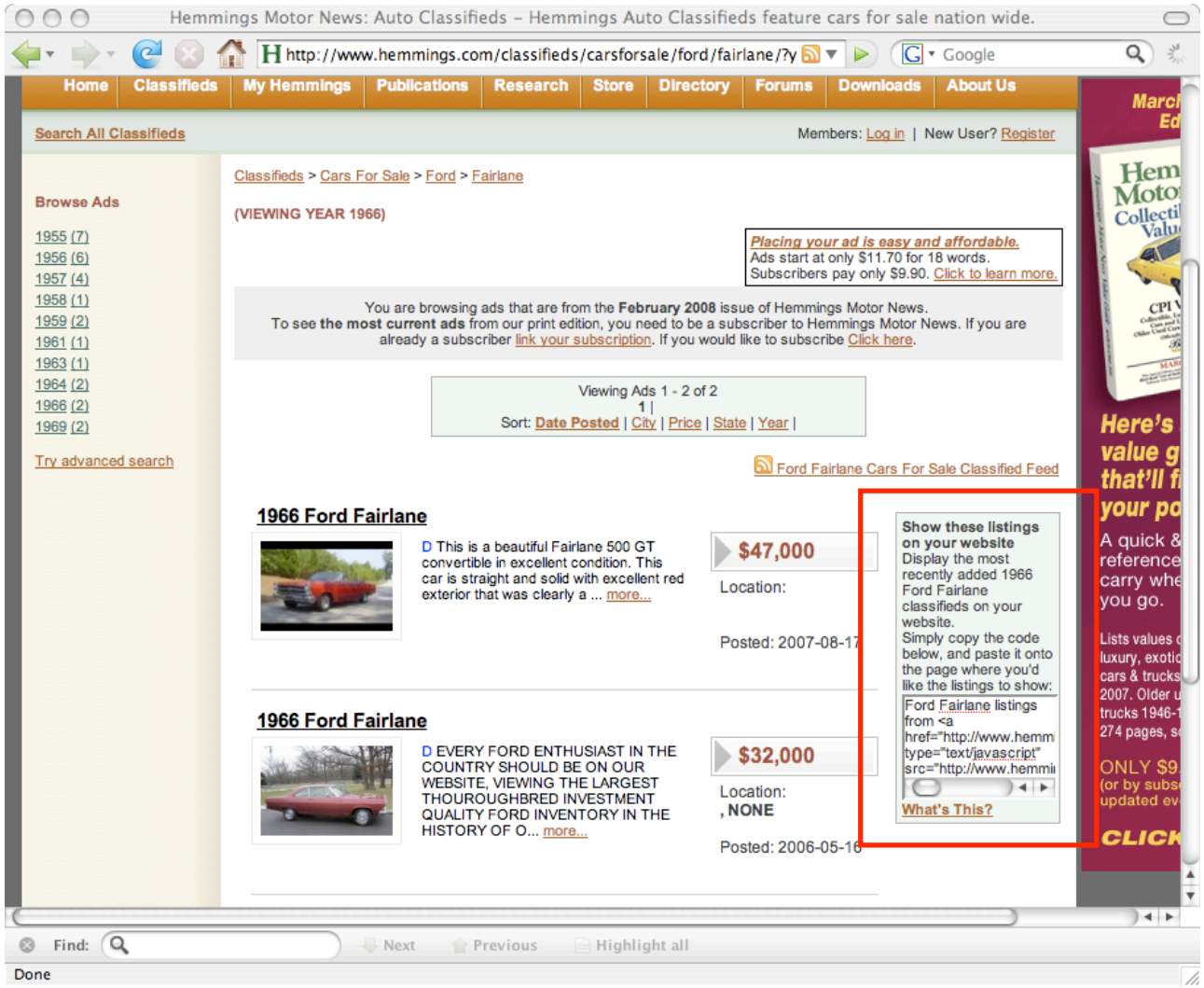
Embedding :: Transparency



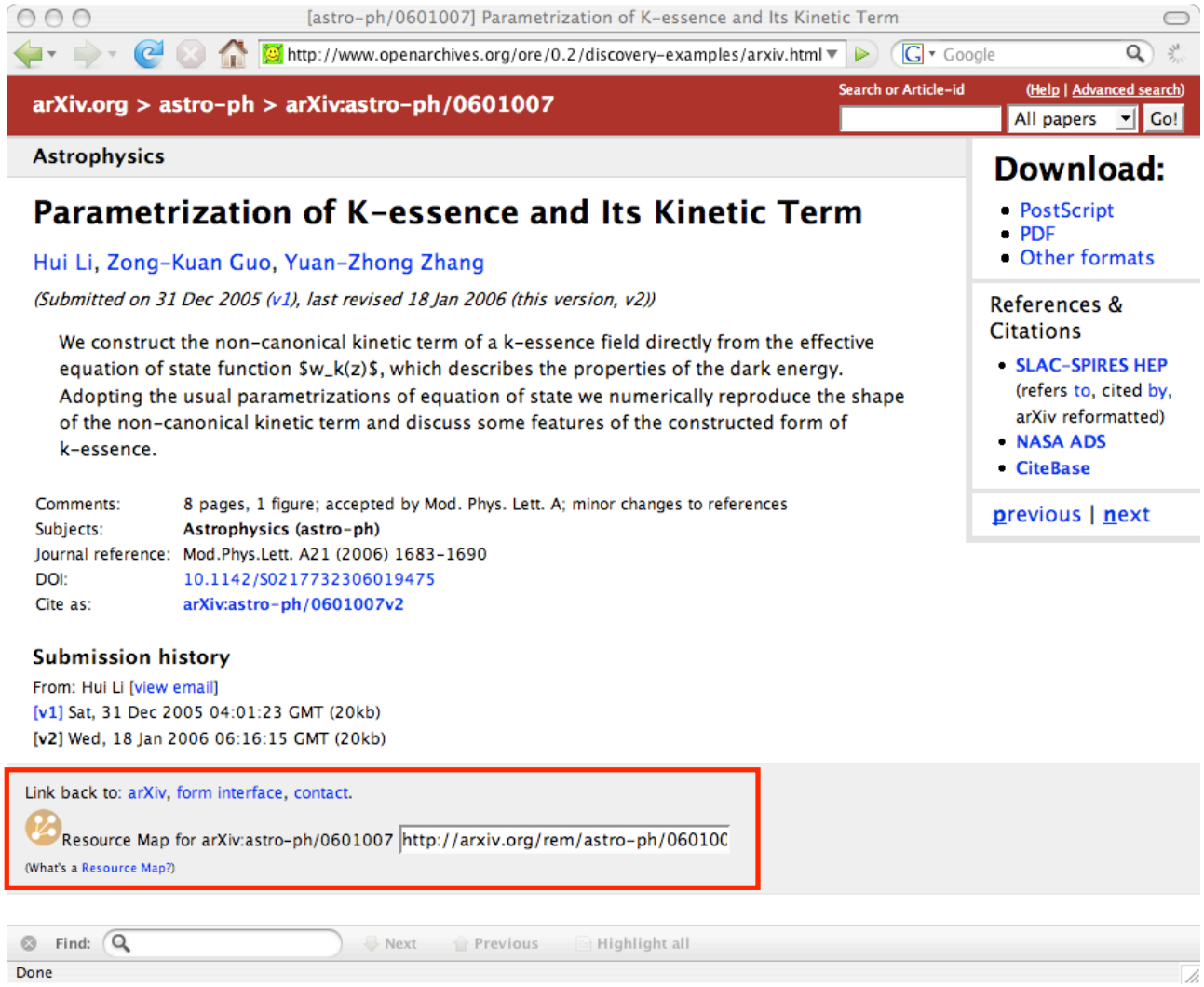
Embedding :: Transparency



Embedding :: Transparency



Embedding :: Transparency



Discovery is a Dirty Job

- Frequently a trade-off between “cleanliness” and “utility”
- Multiple discovery methods, possibly more evolving over time
- Each method has caveats and multiple opportunities to get it wrong
- At least 3 open issues, perhaps more that we have yet to uncover