Metadata and Syndication:
Interoperability and Mashups

CS 431 March 5, 2008
Carl Lagoze - Cornell University
Mashups

• **Combining data from several web sources**
  - Treating the web as a database rather than a document store
• **Post-processing that data**
• **Presenting the processed data**
Combining Data from Multiple Sources

Web data source:
- KML file
- GPX file
- Custom XML file

Map provider:
- Google Map
- Yahoo!Map
- MSN Local
- Ask.com
- Custom maps

Others Web sources
Ex: JSON formatted data
- Non-standard format

Proxy server
Combining Data from Multiple Sources

Web Mashup Styles
In-Browser | Server-side

The Web as Global SOA

HTTP(s)
Web Proxy

REST
JSON
XML
SOAP
RSS
ATOM

The Enterprise

SOAP
WS-*
JMS
MOM

Integration Code
Server-Side Mashup

Ajax Application (Browser)

Source: http://web2.wg2.com
Other types of mashups
What lies underneath?

• *Getting* heterogeneous systems to work together
• *Providing* the user with a seamless information experience
• *Allow* parameterization and interactive experience
  - *AJAX*
INTEROPERABILITY
Dimensions of **Interoperability**

- **Syntax**
  - XML
- **Semantics**
  - XML Schema
  - RDF/RDFS
- **Vocabularies/Ontologies**
  - Dublin Core
  - Simple Knowledge Organisation System (SKOS)
  - OWL
- **Content models**
  - METS
  - FEDORA
  - DIDL
  - ORE
Contrast to Distributed Systems

• Distributed systems
  - Collections of components at different sites that are carefully designed to work with each other

• Heterogeneous or federated systems
  - Cooperating systems in which individual components are designed or operated autonomously
Base Interoperability: web interoperability (HTTP, HTML)

Crawling and Automated Processing (indexing)

“central” index
Crawlers and internet history

- 1991: HTTP
- 1992: 26 servers
- 1993: 60+ servers; self-register; archie
- 1994 (early) - first crawlers
- 1996 - search engines abound
- 1998 - focused crawling
- 1999 - web graph studies
- Current - personalized focused
Metadata aggregation and harvesting

• **Crawling is not always appropriate**
  - rights issues
  - focused targets
  - firewalls
  - deep web
  - Its not all text

• **Other applications than search**
  - Current awareness
  - Preservation
  - Summarization
  - Complex/compound object structure (browsing, etc.)
The general model
Syndication - RSS and Atom

- Format to expose news and content of news-like sites
  - Wired
  - Slashdot
  - Weblogs
- "News" has very wide meaning
  - Any dynamic content that can be broken down into discrete items
    - Wiki changes
    - CVS checkins
- Roles
  - Provider **syndicates** by placing an RSS-formated XML file on Web
  - Aggregator runs RSS-aware program to check feeds for changes
RSS History

• Original design (0.90) for Netscape for building portals of headlines to news sites
  - Loosely RDF based
• Simplified for 0.91 dropping RDF connections
• RDF branch was continued with namespaces and extensibility in RSS 1.0
• Non-RDF branch continued to 2.0 release
• Alternately called:
  - Rich Site Summary
  - RDF Site Summary
  - Really Simple Syndication
RSS is in wide use

- All sorts of origins
  - News
  - Blogs
  - Corporate sites
  - Libraries
  - Commercial
RSS components

- **Channel**
  - single tag that encloses the main body of the RSS document
  - Contains metadata about the channel - **title**, **link**, **description**, **language**, **image**

- **Item**
  - Channel may contain multiple items
  - Each item is a “story”
  - Contains metadata about the story (**title**, **description**, etc.) and possible **link** to the story
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
    <title>NYT > Home Page</title>
    <language>en-us</language>
    <copyright>Copyright 2007 The New York Times Company</copyright>
    <lastBuildDate>Tue, 27 Feb 2007 16:05:01 EST</lastBuildDate>
    <image>
      <title>NYT > Home Page</title>
      <link>http://www.nytimes.com/index.html</link>
    </image>
    <item>
      <title>Wall Street Plummets After Chinese Stocks Take a Big Hit</title>
      <description>Stocks plunged in New York today after a sell-off in China rattled markets worldwide.</description>
      <author>Jeremy W. Peters and David Barboza</author>
      <pubDate>Tue, 27 Feb 2007 15:55:12 EDT</pubDate>
    </item>
    <item>
      <title>Cheney Unhurt After Bombing in Afghanistan</title>
      <description>A suicide bomber blew himself up outside the U.S. base at Bagram while Vice President Dick Cheney was inside. The Taliban claimed responsibility and said Mr. Cheney was the target.</description>
      <author>Abedul Waheed Wafo</author>
      <pubDate>Tue, 27 Feb 2007 15:39:24 EDT</pubDate>
    </item>
  </channel>
</rss>
RSS 2.0 Example
- Namespaces

```xml
<?xml version="1.0" encoding="iso-8859-1"?>
<rss version="2.0" xmlns:photo="http://www.pheed.com/pheed/"
     xmlns:dc="http://purl.org/dc/elements/1.1"/>
  <channel>
    <title>Natural Landscape Photographs</title>
    <link>http://www.photomark.com/cgi-bin/set.cgi?set_id=7</link>
    <description>A few natural landscape photographs.</description>
    <language>en-us</language>
    <item>
      <title>Windmill Farm with Cloud</title>
      <link>http://www.photomark.com/cgi-bin/set.cgi?set_id=7&amp;n=0</link>
      <description>Windmill Farm at dusk with lenticular cloud, Wyoming</description>
      <category>In progress</category>
      <dc:creator>Mark Meyer</dc:creator>
      <dc:rights>Copyright 2001 Mark Meyer</dc:rights>
      <dc:coverage>Wyoming</dc:coverage>
      <dc:format>35mm Transparency</dc:format>
      <dc:subject>Windmill farm lenticular cloud</dc:subject>
      <photo:thumbnail>www.photomark.com/webpix/th/Windmillsa.jpg</photo:thumbnail>
    </item>
    <item>
      <title>The Racetrack Playa</title>
      <link>http://www.photomark.com/cgi-bin/set.cgi?set_id=7&amp;n=1</link>
      <description>The Racetrack Playa</description>
      <category>In progress</category>
      <dc:creator>Mark Meyer</dc:creator>
      <dc:rights>Copyright 2003 Mark Meyer</dc:rights>
      <dc:coverage>Death Valley National Park, California</dc:coverage>
      <dc:format>4x5 Transparency</dc:format>
      <dc:subject>dry desert cracks</dc:subject>
      <photo:thumbnail>http://www.photomark.com/webpix/th/racetrack.jpg</photo:thumbnail>
    </item>
  </channel>
</rss>
```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns="http://purl.org/rss/1.0/
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 >
 <channel rdf:about="http://example.com/news.rss">
   <title>Example Channel</title>
   <link>http://example.com/</link>
   <description>My example channel</description>
   <items>
     <rdf:Seq>
     </rdf:Seq>
   </items>
 </channel>
</rdf:RDF>
Atom

• Attempt to rationalize RSS 1.x, 2.x divergence
• Encoding is up-to-date with current XML standards
  - namespaces
  - Schema
• Robust content model
  - Distinguishes between metadata and content (plain text, HTML, base-64 binary)
• Well-defined extensibility model
• IETF FRC 4287
  - http://www.ietf.org/rfc/rfc4287
Simple Atom Feed

```xml
<?xml version="1.0" encoding="UTF-8"?>
<feed xmlns="http://www.w3.org/2005/Atom"
      xml:lang="en"
      xml:base="http://www.example.org">
  <id>http://www.example.org/myfeed</id>
  <title>My Simple Feed</title>
  <updated>2005-07-15T12:00:00Z</updated>
  <link href="/blog" />
  <link rel="self" href="/myfeed" />

  <entry>
    <id>http://www.example.org/entries/1</id>
    <title>A simple blog entry</title>
    <link href="/blog/2005/07/1" />
    <updated>2005-07-15T12:00:00Z</updated>
    <summary>This is a simple blog entry</summary>
  </entry>

  <entry>
    <id>http://www.example.org/entries/2</id>
    <title/>
    <link href="/blog/2005/07/2" />
    <updated>2005-07-15T12:00:00Z</updated>
    <summary>This is simple blog entry without a title</summary>
  </entry>

</feed>
```
Atom with namespaces

<?xml version="1.0" encoding="UTF-8"?>
<feed xmlns="http://www.w3.org/2005/Atom" xml:lang="en" xmlns:foaf="http://xmlns.com/foaf/0.1"
      xmlns:wot="http://purl.org/wot"
      xmlns:vcard="http://purl.org/vcard/ns#">
  <id>http://www.example.org/myfeed</id>
  <title>My Simple Feed</title>
  <updated>2005-07-15T12:00:00Z</updated>
  <dc:creator>
    <name>James M Snell</name>
    <foaf:homepage rdf:resource="/blog"/>
    <foaf:img rdf:resource="/myopic.png"/>
  </dc:creator>
  <dc:contributor>
    <name>Jane Doe</name>
    <foaf:homepage rdf:resource="/janesblog"/>
    <foaf:img rdf:resource="/janespic.png"/>
  </dc:contributor>
  <link href="/blog"/>
  <link rel="self" href="/myfeed"/>
  <entry>
    <id>http://www.example.org/entries/1</id>
    <title>A simple blog entry</title>
    <link href="/blog/2005/07/1"/>
    <updated>2005-07-15T12:00:00Z</updated>
    <summary>This is a simple blog entry</summary>
  </entry>
  <entry>
    <id>http://www.example.org/entries/2</id>
    <title/>
    <link href="/blog/2005/07/2"/>
    <updated>2005-07-15T12:00:00Z</updated>
    <summary>This is simple blog entry without a title</summary>
  </entry>
</feed>
Atom Enclosures and Content Support (podcast)

<feed xmlns="http://www.w3.org/2005/Atom">
  <id>http://www.example.org/myfeed</id>
  <title>My Podcast Feed</title>
  <updated>2005-07-15T12:00:00Z</updated>
  <author>
    <name>James M Snell</name>
  </author>
  <link href="http://example.org" />
  <link rel="self" href="http://example.org/myfeed" />
  <entry>
    <id>http://www.example.org/entries/1</id>
    <title>Atom 1.0</title>
    <updated>2005-07-15T12:00:00Z</updated>
    <link href="http://www.example.org/entries/1" />
    <summary>An overview of Atom 1.0</summary>
    <link rel="enclosure"
      type="audio/mpeg"
      title="MP3"
      href="http://www.example.org/myaudiofile.mp3"
      length="1234" />
    <link rel="enclosure"
      type="application/x-bittorrent"
      title="BitTorrent"
      href="http://www.example.org/myaudiofile.torrent"
      length="1234" />
    <content type="xhtml">
      <div xmlns="http://www.w3.org/1999/xhtml">
        <h1>Show Notes</h1>
        <ul>
          <li>00:01:00 -- Introduction</li>
          <li>00:15:00 -- Talking about Atom 1.0</li>
          <li>00:30:00 -- Wrapping up</li>
        </ul>
      </div>
    </content>
  </entry>
</feed>
Automated discovery of RSS/ATOM feeds
What RSS doesn’t have

- Notion of a “collection” - corpus of documents that persist
- Technique for selectively requesting metadata from parts of the collection
- Notion of multiple descriptive types

- These things are important for more “library-like” corpora, e.g., museums, libraries, repositories
The Open Archives Initiative (OAI) and the Protocol for Metadata Harvesting (OAI-PMH)
OAI-PMH

⇒ PMH -> Protocol for Metadata Harvesting [http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm](http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm)
• Simple protocol, just 6 verbs
• Designed to allow harvesting of any XML (meta)data (schema described)
• For batch-mode not interactive use
OAI for discovery

User

R1

R2

R3

R4

Information islands
OAI for discovery

Service layer

User

Search service

Metadata harvested by service

R1
R2
R3
R4
OAI-based Search

OAI-PMH Data Model

item has identifier

all available metadata about this sculpture

resource

item

records

record has identifier + metadata format + datestamp

Dublin Core metadata

MARC21 metadata

branding metadata
Identifiers

• Items have identifiers (all records of same item share identifier)
• Identifiers must have URI syntax identifiers must be assumed to be local to the repository
• Complete identification of a record is baseURL+identifier+metadataPrefix+datestamp
## OAI-PMH verbs

<table>
<thead>
<tr>
<th>Verb</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify</td>
<td>description of archive</td>
</tr>
<tr>
<td>ListMetadataFormats</td>
<td>metadata formats supported by archive</td>
</tr>
<tr>
<td>ListSets</td>
<td>sets defined by archive</td>
</tr>
<tr>
<td>ListIdentifiers</td>
<td>OAI unique ids contained in archive</td>
</tr>
<tr>
<td>ListRecords</td>
<td>listing of N records</td>
</tr>
<tr>
<td>GetRecord</td>
<td>listing of a single record</td>
</tr>
</tbody>
</table>

Most verbs take arguments: dates, sets, ids, metadata formats and resumption token (for flow control)
OAI-PMH and HTTP

- OAI-PMH uses HTTP as transport
  - Encoding OAI-PMH in GET
    - http://baseURL?verb=<verb>&arg1=<arg1Val>...
    - Example: http://an.oa.org/OAIscript?
      verb=GetRecord&
      identifier=oai:arXiv.org:hep-th/9901001&
      metadataPrefix=oai_dc

- Error handling
  - all OK at HTTP level? => 200 OK
  - something wrong at OAI-PMH level? => OAI-PMH error (e.g. badVerb)

- HTTP codes 302 (redirect), 503 (retry-after), etc. still available to implementers, but do not represent OAI-PMH events
OAI and Metadata Formats

• Protocol based on the notion that a record can be described in multiple metadata formats
• Dublin Core is required for “interoperability”
OAI-PMH Responses

- All defined by one schema
  - http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd

- Generic Structure (Header and Body)

```xml
<?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>2007-02-28T12:50:01Z</responseDate>
  <request verb="ListRecords" metadataPrefix="oai_dc" set="mussm"
    >http://memory.loc.gov/cgi-bin/oai2_0</request>
  <ListRecords>[5168 lines]</ListRecords>
</OAI-PMH>
```
Generic Record Structure

<record>
<header>
<identifier>oai:loc.gov:loc.music/sm1819.360010</identifier>
<datestamp>2005-11-21T17:08:59Z</datestamp>
<setSpec>mussm</setSpec>
</header>
<metadata>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
<xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>The hunter's horn, a new sporting cavatina</dc:title>
<dc:creator>Philipps, T.</dc:creator>
<dc:subject>Cavatina</dc:subject>
<dc:subject>Songs with piano</dc:subject>
<dc:description>In bound volumes: Copyright Deposits 1820-1860</dc:description>
<dc:publisher>New York: Geib and Co</dc:publisher>
<dc:date>1819</dc:date>
<dc:type>text</dc:type>
<dc:type>musical notation</dc:type>
<dc:identifier>http://hdl.loc.gov/loc.music/sm1819.360010</dc:identifier>
<dc:language>eng</dc:language>
</oai_dc:dc>
</metadata>
</record>
PAI-PMH Requests

- http://memory.loc.gov/cgi-bin/oai2_0?verb=ListMetadataFormats
- http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&metadataPrefix=oai_dc
- http://memory.loc.gov/cgi-bin/oai2_0?verb=ListRecords&metadataPrefix=oai_marc
Selective Harvesting

- RSS is mainly a “tail” format
- OAI-PMH is more “grep” like
- Two “selectors” for harvesting
  - Date
  - Set
- Why not general search?
  - Out of scope
  - Not low-barrier
  - Difficulty in achieving consensus
Datestamps

• All dates/times are UTC, encoded in ISO8601, Z notation:
  1957-03-20T20:30:00Z

• Datestamps may be either fill date/time as above or date only (YYYY-MM-DD). Must be consistent over whole repository, 'granularity' specified in Identify response.

• Earlier version of the protocol specified “local time” which caused lots of misunderstandings. Not good for global interoperability!
Sets

- Simple notion of grouping at the item level to support selective harvesting
  - Hierarchical set structure
  - Multiple set membership permitted
  - E.g: repo has sets A, A:B, A:B:C, D, D:E, D:F
    If item1 is in A:B then it is in A
    If item2 is in D:E then it is in D, may also be in D:F
    Item3 may be in no sets at all

http://memory.loc.gov/cgi-bin/oai2_0?verb=ListSets
Selective Harvesting Request

Harvesting strategy

- **Issue Identify request**
  - Check all as expected (validate, version, baseURL, granularity, comporession...)
- **Check sets/metadata formats as necessary (ListSets, ListMetadataFormats)**
- Do harvest, initial complete harvest done with no from and to parameters
- Subsequent incremental harvests start from datastamp that is responseDate of last response
OAI-PMH - Has it worked?

• Of course, yes...
  - Very wide deployment
  - “millions and millions of records served”
  - Incorporated into commercial systems

• But....
  - NSDL experience has shown “low barrier” is not always true
    • XML is hard
  - Incremental harvesting model is full of holes