Using metadata for interoperability

CS 431 February 28, 2007 Carl Lagoze – Cornell University What is the problem?

- Getting heterogeneous systems to work together
- Providing the user with a seamless information experience
- What services do you want to provide?
 - Search and access?
 - More?
 - information access
 - authorization and authentication
 - integrity and reliability
 - Reuse
- How much human intervention?
- Level of perfection?

Why is it hard

- Differences in...
 - hardware
 - applications
 - design patterns
 - language
 - culture
 - laws
 - policies
 - human behaviors

Interoperability is multidimensional

- Syntax
 - XML
- Semantics
 - RDF/RDFS/OWL
- Vocabularies/Ontologies
 - Dublin Core/ABC/CIDOC-CRM
- Search and discovery
 - Z39.50
 - SDLIP
 - ZING
- Protocols
 - Dienst
 - OAI-PMH
- Information models
 - METS
 - FEDORA
 - DIDL

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

Web Search Strategies - Crawling and Automated Indexing



Definition

```
Spider = robot = crawler
```

Crawlers are computer programs that roam the Web with the goal of automating specific tasks related to the Web.

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
- 2002 use for digital libraries (focused crawling)

Metadata aggregation and harvesting

- Crawling is not always appropriate
 - rights issues
 - focused targets
 - firewalls
 - deep web
- 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

Simple RSS 2.0 Example

<?xml version="1.0" encoding="UTF-8"?> <rss version="2.0"> <channel> <title>NYT > Home Page</title> k>http://www.nytimes.com/index.html?partner=rssnyt</link> <description>New York Times > Breaking News, World News & amp; Multimedia</description> <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> <url>http://graphics.nytimes.com/images/section/NytSectionHeader.gif</url> k>http://www.nytimes.com/index.html</link> </image> <item> <title>Wall Street Plummets After Chinese Stocks Take a Big Hit</title> link>http://www.nytimes.com/2007/02/28/business/28stox.web.html?ex=1330318800&en=43b57471d <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> <guid isPermaLink="false">http://www.nvtimes.com/2007/02/28/business/28stox.web.html</guid> <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>ABDUL WAHEED WAFA</author> <guid isPermaLink="false">http://www.nvtimes.com/2007/02/27/world/asia/27cnd-chenev.html</guid> <pubDate>Tue, 27 Feb 2007 15:39:24 EDT</pubDate> </item>

RSS 2.0 Example - Namespaces

<?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>

k>http://www.photo-mark.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>

k>http://www.photomark.com/cgi-bin/set.cgi?set_id=7&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:imgsrc> http://www.photo-mark.com/webpix/ds/Windmillsa.jpg </photo:imgsrc>
<photo:thumbnail> www.photo-mark.com/webpix/tn/Windmillsa.jpg </photo:thumbnail>

</item>

<item>

<title>The Racetrack Playa</title>

k> http://www.photo-mark.com/cgi-bin/set.cgi?set_id=7&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:imgsrc> http://www.photo-mark.com/webpix/ds/racetrack.jpg </photo:imgsrc>
<photo:thumbnail> http://www.photo-mark.com/webpix/tn/racetrack.jpg </photo:thumbnail>

</item> </channel> </rss>

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 version="1.0" encoding="UTF-8"?> <feed xmIns="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> k href="/blog" /> k 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 xmins="http://www.w3.org/2005/Atom" xml:lang="en" xmins:foaf="http://xmins.com/foaf/0.1"
  xml:base="http://www.example.org" xmlns:dc="http://purl.org/dc/elements/1.1/">
  <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"/>
    <foaftimg rdf:resource="/mypic.png"/>
  </dc:creator>
  <dc:contributor>
    <name>Jane Doe</name>
    <foaf:homepage rdf:resource="/janesblog"/>
    <foaf:image rdf:resource="/janespic.png"/>
  </dc:contributor>
  k href="/blog"/>
  k rel="self" href="/myfeed"/>
  <entry>
    <id>http://www.example.org/entries/1</id>
    <title>A simple blog entry</title>
    k 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/>
    k 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)

<?xml version="1.0" encoding="UTF-8"?> <feed,xmIns="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" /> k rel="self" href="http://example.org/myfeed" /> <entrv> <id>http://www.example.org/entries/1</id> <title>Atom 1.0</title> <updated>2005-07-15T12:00:00Z</updated> k 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 xmIns="http://www.w3.org/1999/xhtml"> <h1>Show Notes</h1> <u>00:01:00 -- Introduction 00:15:00 -- Talking about Atom 1.0 li>00:30:00 -- Wrapping up </div> </content> </entry> </feed>.

Automated discovery of RSS/ATOM feeds

<!-- feed autodiscovery links -->

- k rel="alternate" type="application/atom+xml"
- title="XML.com Articles and Weblogs" href="http://www.oreillynet.com/pub/feed/20" />

k rel="alternate" type="application/rdf+xml"

title="XML.com Articles and Weblogs" href="http://www.oreillynet.com/pub/feed/20?format=rss1" /> <link rel="alternate" type="application/rss+xml"

title="XML.com Articles and Weblogs" href="http://www.oreillynet.com/pub/feed/20?format=rss2" />

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, *institutional repositories*

The Open Archives Initiative (OAI) and the Protocol for Metadata Harvesting (OAI-PMH)



OAI-PMH

- PMH -> Protocol for Metadata Harvesting <u>http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm</u>
- 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



OAI for discovery



Metadata harvested by service

OAI for XYZ



OAI-PMH Data Model



record has identifier + metadata format + datestamp

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

	Verb	Function
metadata about the repository	Identify	description of archive
	ListMetadataFormats	metadata formats supported by archive
	ListSets	sets defined by archive
harvesting verbs	ListIdentifiers	OAI unique ids contained in archive
	ListRecords	listing of N records
	GetRecord	listing of a single record

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
 - <u>http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd</u>"
- Generic Structure (Header and Body)

<?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"> %ttp://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"> %ttp://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] </OAI-PMH>

Generic Record Structure

<record> <header> <identifier>oai:lcoa1.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>

Identify: Information about repository

http://memory.loc.gov/cgi-bin/oai2_0?verb=Identify

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmIns="http://www.openarchives.org/OAI/2.0/"
     xmIns: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>2007-02-28T12:18:51Z</responseDate>
<request verb="Identify">http://memory.loc.gov/cgi-bin/oai2_0</request>
 <Identify>
 <repositoryName>Library of Congress Open Archive Initiative Repository 1</repositoryName>
 <baseURL>http://memory.loc.gov/cgi-bin/oai2_0</baseURL>
 <protocolVersion>2.0</protocolVersion>
 <adminEmail>dwoo&#x0040;loc.gov</adminEmail>
 <adminEmail>caar&#x0040:loc.gov</adminEmail>
 <earliestDatestamp>2002-06-01T00:002</earliestDatestamp>
  <deletedRecord>no</deletedRecord>
 <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
 <description>
  <oai-identifier
    xmIns="http://www.openarchives.org/OAI/2.0/oai-identifier"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai-identifier
     http://www.openarchives.org/OAI/2.0/oai-identifier.xsd">
    <scheme>oai</scheme>
   <repositoryIdentifier>Icoa1.loc.gov</repositoryIdentifier>
   <delimiter>:</delimiter>
   <sampleIdentifier>oai:lcoa1.loc.gov:loc.music/musdi.002</sampleIdentifier>
   </oai-identifier>
 </description>
```

ListMetadataFormats: Available Formats

http://memory.loc.gov/cgi-bin/oai2_0?verb=ListMetadataFormats

<?xml version="1.0" encoding="UTF-8"?> <OAI-PMH xmIns="http://www.openarchives.org/OAI/2.0/" xmIns: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>2007-02-28T12:48:02Z</responseDate> <request verb="ListMetadataFormats">http://memory.loc.gov/cgi-bin/oai2_0</request> <ListMetadataFormats> <metadataFormat> <metadataPrefix>oai dc</metadataPrefix> <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema> <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace> </metadataFormat> <metadataFormat> <metadataPrefix>oai marc</metadataPrefix> <schema>http://www.openarchives.org/OAI/1.1/oai_marc.xsd</schema> <metadataNamespace>http://www.openarchives.org/OAI/1.1/oai_marc</metadataNamespace> </metadataFormat> <metadataFormat> <metadataPrefix>marc21</metadataPrefix> <schema>http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd</schema> <metadataNamespace>http://www.loc.gov/MARC21/slim</metadataNamespace> </metadataFormat> <metadataFormat> <metadataPrefix>mods</metadataPrefix> <schema>http://www.loc.gov/standards/mods/v3/mods-3-0.xsd</schema> <metadataNamespace>http://www.loc.gov/mods/v3</metadataNamespace> </metadataFormat> </ListMetadataFormats> </OAI-PMH>

ListRecords: Retrieve Metadata Records

<?xml version="1.0" encoding="UTF-8"?> <OAI-PMH xmIns="http://www.openarchives.org/OAI/2.0/" xmIns: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</reguest> <ListRecords> <record> <header> <identifier>oai:lcoa1.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> <record> <header> <identifier>oai:lcoa1.loc.gov:loc.music/sm1819.360020</identifier> <datestamp>2005-11-21T17:08:59Z</datestamp> 1.10

http://memory.loc.gov/cgioai2_0?verb=ListRecords&metadataP refix=oai_dc&set=mussm

Error/exception response

http://memory.loc.gov/cgibin/oai2_0?verb=ListRecords&metadataPrefix=WRONG

<?xml version="1.0" encoding="UTF-8"?> <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlnstxsi="http://www.oyenarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"> <responseDate>2007-02-28T13:05:06Z</responseDate> <request verb="ListRecords" metadataPrefix="WRONG">http://memory.loc.gov/cgi-bin/oai2_0</request> <error code="cannotDisseminateFormat">Metadata prefix (WRONG) not supported by this repository.</error> </OAI-PMH>

resumptionToken

- Protocol supports the notion of partial responses in a very simple way: Response includes a 'token' at the which is used to get the next chunk.
- Idempotency of resumptionToken: return same incomplete list when resumptionToken is reissued
 - while no changes occur in the repo: strict
 - while changes occur in the repo: all items with unchanged datestamp
 - optional attributes for the resumptionToken: expirationDate, completeListSize, cursor

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

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