Data, Metadata, and Identifiers

Christopher Jones, Software Engineer
National Center for Ecological Analysis and Synthesis
University of California, Santa Barbara

International Digital Curation Conference
February 27, 2014
Topics
Topics

Everything is an Object
Topics

Everything is an Object

Types of Metadata
Topics

Everything is an Object

Types of Metadata

Packaging
Topics

Everything is an Object
Types of Metadata
Packaging
Identifiers
Everything is an Object

Science Data
Everything is an Object

Science Data
Everything is an Object
Everything is an Object

Data are discreet sets of bytes
Everything is an Object
Everything is an Object

DataONE is inclusive of all data types
Community repositories may focus on specific data types
DataONE facilitates preservation and discovery
Using structured metadata
Types of Metadata
Types of Metadata

High quality metadata

- Promote discovery
- Promote data longevity
- Promote interoperability
Types of Metadata
Types of Metadata

science data
Types of Metadata

System Metadata

- system metadata
- science data
Types of Metadata

System Metadata
Science Metadata

XML

science metadata

science data

system metadata

system metadata
Types of Metadata

System Metadata

Science Metadata

Resource Maps

XML

resource map

- system metadata
- science metadata

system metadata

science metadata

science data
Types of Metadata
Types of Metadata

System Metadata

<table>
<thead>
<tr>
<th>Identifier</th>
<th>doi:10.5063/AA/nceas.144.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>53267894 bytes</td>
</tr>
<tr>
<td>Checksum</td>
<td>SHA1 A3487BCE458 ...</td>
</tr>
<tr>
<td>Date Uploaded</td>
<td>20140201T040802124</td>
</tr>
<tr>
<td>Access Policy</td>
<td>public: read</td>
</tr>
<tr>
<td>Replication Policy</td>
<td>numReplicas: 3</td>
</tr>
<tr>
<td>etc.</td>
<td>...</td>
</tr>
</tbody>
</table>
Types of Metadata

System Metadata

<table>
<thead>
<tr>
<th>Identifier</th>
<th>doi:10.5063/AA/nceas.144.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>53267894 bytes</td>
</tr>
<tr>
<td>Checksum</td>
<td>SHA1 A3487BCE458 ...</td>
</tr>
<tr>
<td>Date Uploaded</td>
<td>20140201T040802124</td>
</tr>
<tr>
<td>Access Policy</td>
<td>public: read</td>
</tr>
<tr>
<td>Replication Policy</td>
<td>numReplicas: 3</td>
</tr>
<tr>
<td>etc.</td>
<td>...</td>
</tr>
</tbody>
</table>

http://mule1.dataone.org/ArchitectureDocs-current/apis/Types.html#Types.SystemMetadata
Types of Metadata
Types of Metadata

Science Metadata

<table>
<thead>
<tr>
<th>Title</th>
<th>Decline in Carbon Assimilation of Forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>Ross McMurtrie</td>
</tr>
<tr>
<td>Abstract</td>
<td>... information on the location, management history, N inputs, N losses, soil, water, ...</td>
</tr>
<tr>
<td>Time Range</td>
<td>1998-01-01 : 1998-12-31</td>
</tr>
<tr>
<td>Spatial Range</td>
<td>N -10.5, S -39.375, E ...</td>
</tr>
<tr>
<td>Methods</td>
<td>... were synthesized from multiple data sources including ...</td>
</tr>
<tr>
<td>etc.</td>
<td>...</td>
</tr>
</tbody>
</table>
Types of Metadata
Types of Metadata

Science Metadata

- Ecological Metadata Language
- FGDC CSDGM
- FGDC Biological Data Profile
- ESRI FGDC Profile
- ISO 19115
- Dryad Metadata Profile
- Dublin Core
Types of Metadata

Science Metadata

- Ecological Metadata Language
- FGDC CSDGM
- FGDC Biological Data Profile
- ESRI FGDC Profile
- ISO 19115
- Dryad Metadata Profile
- Dublin Core

http://mule1.dataone.org/ArchitectureDocs-current/design/SearchMetadata.html
Types of Metadata
Types of Metadata

Resource Maps

• Open Archives Initiative
• Object Reuse and Exchange
• Defines an ‘aggregation’
• Associates science data and science metadata
• Highly extensible
• RDF/XML syntax
Types of Metadata

Resource Maps
Types of Metadata

Resource Maps

- Resource Maps

- XML

- System metadata

- Science metadata

- Science data
Types of Metadata

Resource Maps

resource map

science metadata  science data
Types of Metadata

Resource Maps

resource map

aggregates

science metadata

science data
Types of Metadata

Resource Maps

- science metadata
- science data

resource map

aggregates

XML
Types of Metadata

Resource Maps

XML

resource map

science metadata

aggregates
Types of Metadata

Resource Maps

- resource map
  - documents
  - science metadata
  - science data
Types of Metadata

Resource Maps

resource map

documents

science metadata

science data
Types of Metadata

Resource Maps

resource map

isDocumentedBy

science metadata

science data
Types of Metadata

Resource Maps

resource map

isDocumentedBy

science metadata

science data
Types of Metadata

Resource Maps

resource map

isDocumentedBy

science metadata

science data
Types of Metadata

Resource Maps

resource map

isDocumentedBy

science metadata

science data
Packaging

Data Package
Packaging

Data Package

http://mule1.dataone.org/ArchitectureDocs-current/design/DataPackage.html
Packaging
Packaging

Resource maps express collections independently of metadata standards
Identifiers
Identifiers

But how do we reference objects?
Identifiers

Data Package

identifier: cjones.1.1.rdf

identifier: cjones.2.1.csv

identifier: cjones.3.1.xml
Identifiers

Data Package

resource map

- system metadata
- system metadata

science metadata

science data

identifier: cjones.1.1.rdf

identifier: cjones.2.1.csv

identifier: cjones.3.1.xml
Identifiers

Limited to 800 characters, no whitespace
Identifiers

EZID
Long-term identifiers made easy

doi

URI
URL URN
Identifiers

Support for arbitrary schemes
Identifiers

Support for arbitrary schemes

- lake-mendota.20130108
- lake-mendota.2013.1.csv
- ark:/13030/m5qj7grq/1/lake-mendota.2013.1.csv
- http://dx.doi.org/10.6073/AA/lake-mendota.2013.1
- 门多塔.2013.1
- 68AF6874-6548-4DC7-B798-81B41BB97851
Identifiers
The underlying bytes of an object referenced by a given identifier must not change.
Identifiers
Identifiers

GLEON

lake-mendota.2013.1

US-LTER
Identifiers

GLEON

lake-mendota.2013.1

US-LTER

two replicas registered
Identifiers
Identifiers

v1 API: objects are immutable
v2 API (in planning): objects may be mutable
Identifiers
Identifiers

immutable: for a given identifier, the underlying bytes never change