

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



.edu .gov
.org .com
repositories

| | | | | | | |
|--|--|--|--|--|--|--|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

<meta>



Community repositories may
focus on specific data types



.edu .gov
.org .com
repositories

| | | | | | | |
|--|--|--|--|--|--|--|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

<meta>



.edu .gov
.org .com
repositories

| | | | | | | |
|--|--|--|--|--|--|--|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

<meta>

DataONE facilitates
preservation and discovery



.edu .gov
.org .com
repositories

| | | | | | | |
|--|--|--|--|--|--|--|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

<meta>



.edu .gov
.org .com
repositories

| | | | | | |
|--|--|--|--|--|--|
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |

<meta>



Using structured metadata

Types of Metadata



Types of Metadata



High quality metadata

- Promote discovery
- Promote data longevity
- Promote interoperability

Types of Metadata



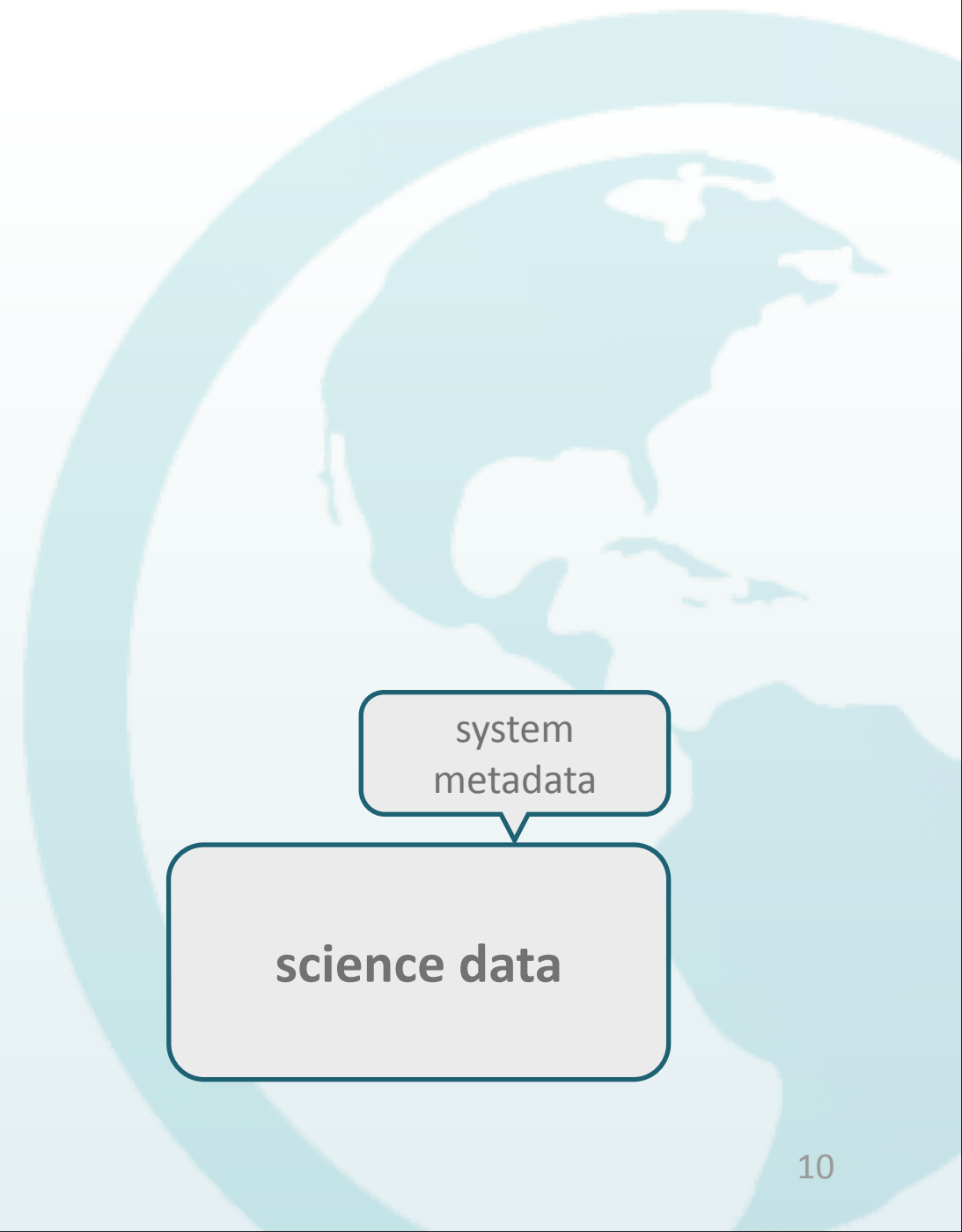
Types of Metadata



science data

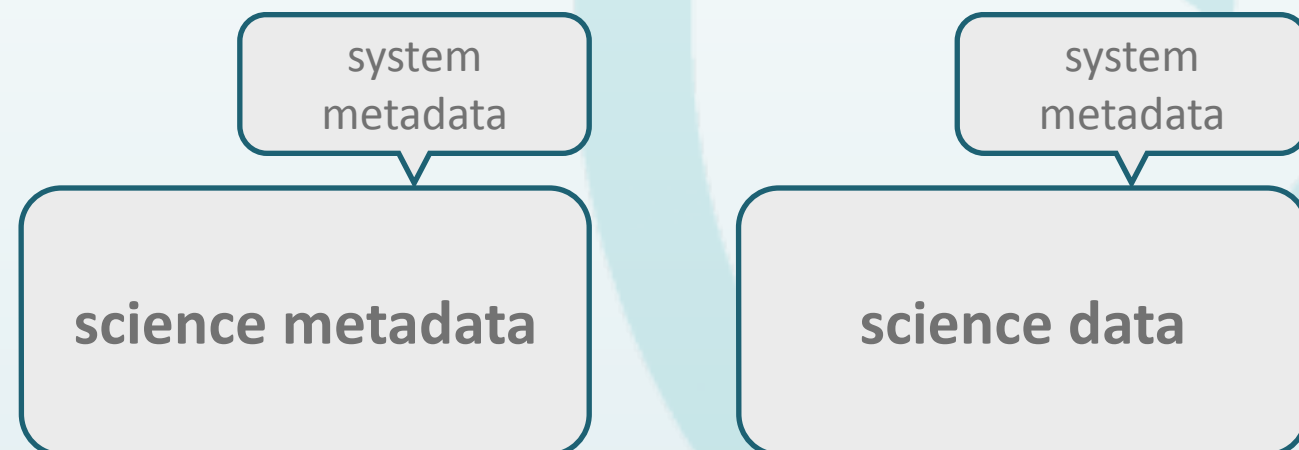
Types of Metadata

System Metadata



Types of Metadata

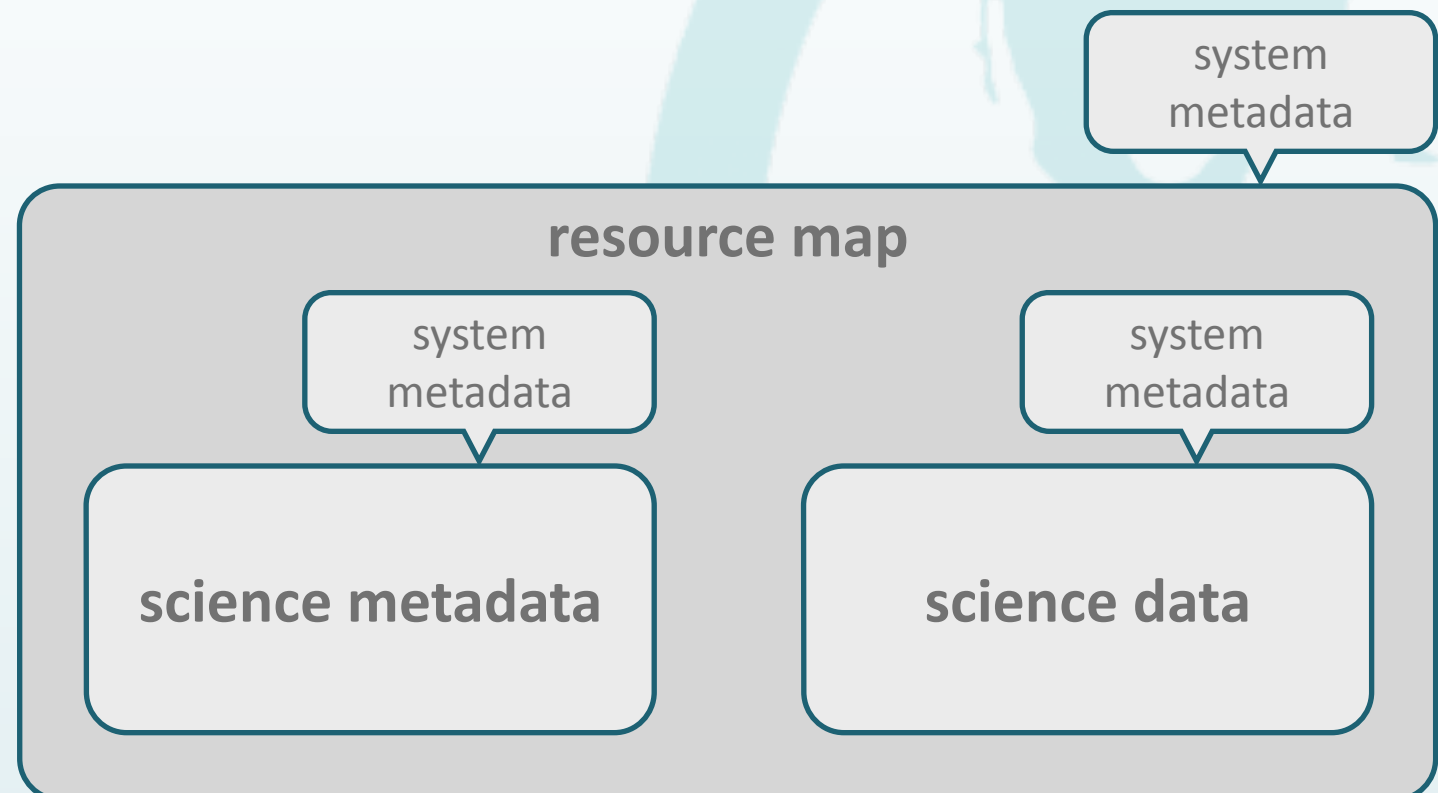
System Metadata
Science Metadata



Types of Metadata



System Metadata
Science Metadata
Resource Maps



Types of Metadata



Types of Metadata

System Metadata



| | |
|--------------------|---|
| Identifier | <code>doi:10.5063/AA/nceas.144.2</code> |
| Size | <code>53267894 bytes</code> |
| Checksum | <code>SHA1 A3487BCE458 ...</code> |
| Date Uploaded | <code>20140201T040802124</code> |
| Access Policy | <code>public: read</code> |
| Replication Policy | <code>numReplicas: 3</code> |
| etc. | <code>...</code> |

Types of Metadata

System Metadata



| | |
|--------------------|---|
| Identifier | <code>doi:10.5063/AA/nceas.144.2</code> |
| Size | <code>53267894 bytes</code> |
| Checksum | <code>SHA1 A3487BCE458 ...</code> |
| Date Uploaded | <code>20140201T040802124</code> |
| Access Policy | <code>public: read</code> |
| Replication Policy | <code>numReplicas: 3</code> |
| etc. | <code>...</code> |

<http://mule1.dataone.org/ArchitectureDocs-current/apis/Types.html#Types.SystemMetadata>

Types of Metadata



Types of Metadata

Science Metadata



| | |
|---------------|---|
| Title | Decline in Carbon Assimilation of Forests |
| Creator | Ross McMurtrie |
| Abstract | ... information on the location, management history, N inputs, N losses, soil, water, ... |
| Time Range | 1998-01-01 : 1998-12-31 |
| Spatial Range | N -10.5, S -39.375, E ... |
| Methods | ... were synthesized from multiple data sources including ... |
| etc. | ... |

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

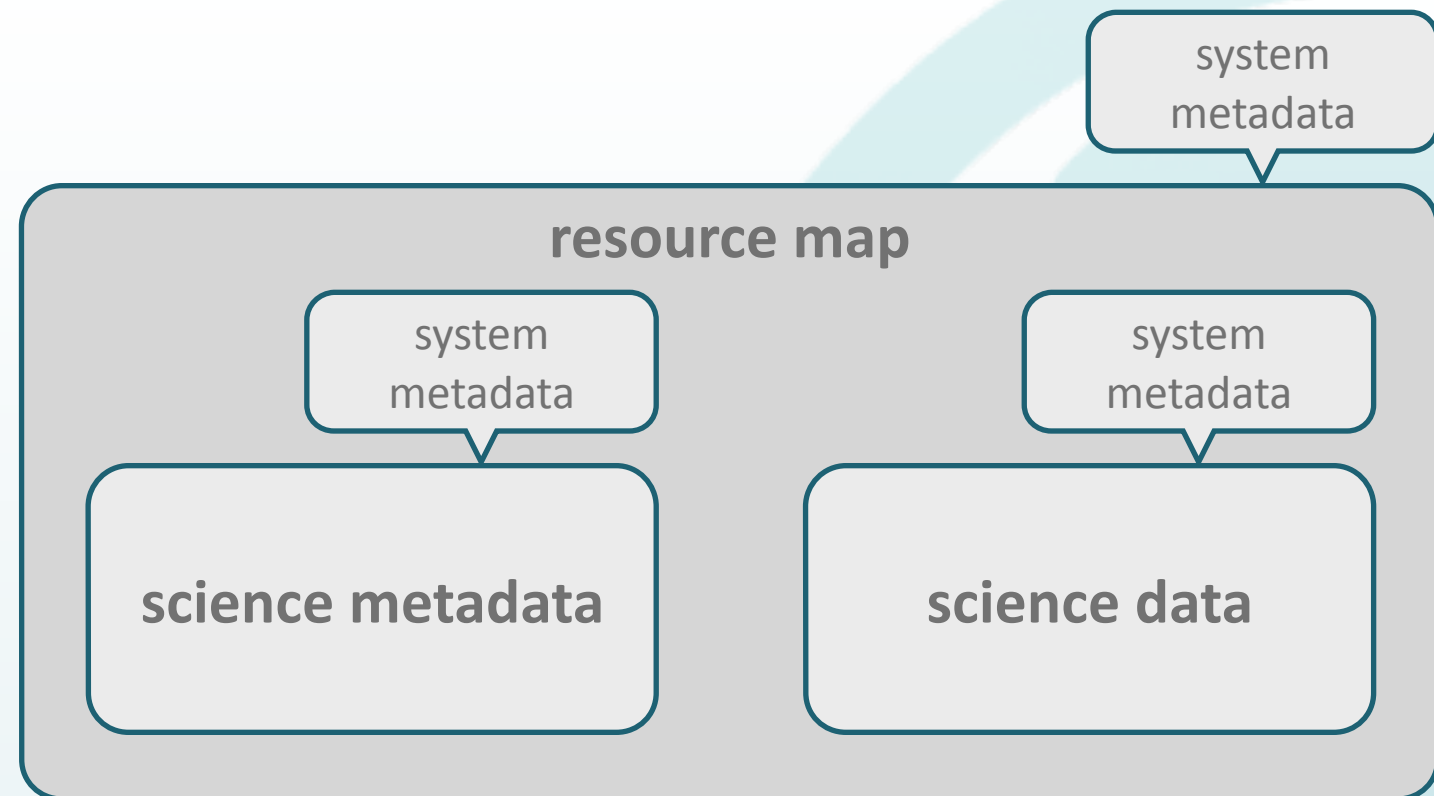
Types of Metadata

Resource Maps



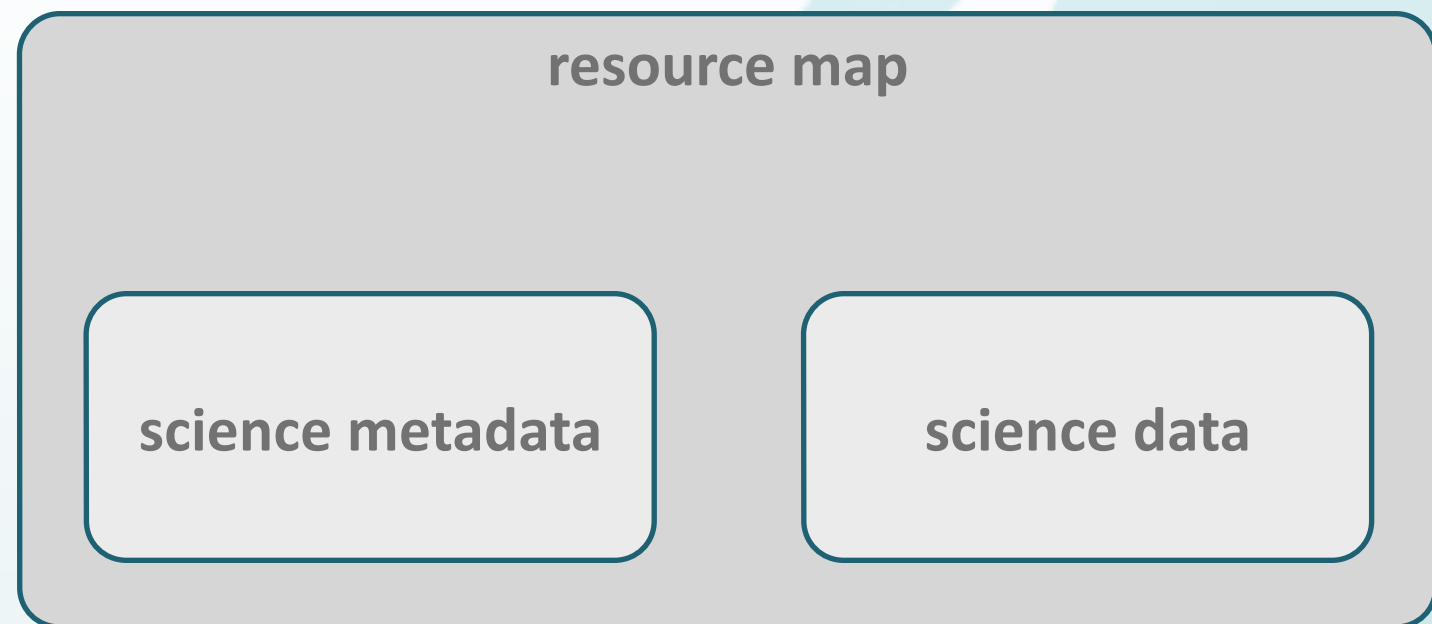
Types of Metadata

Resource Maps



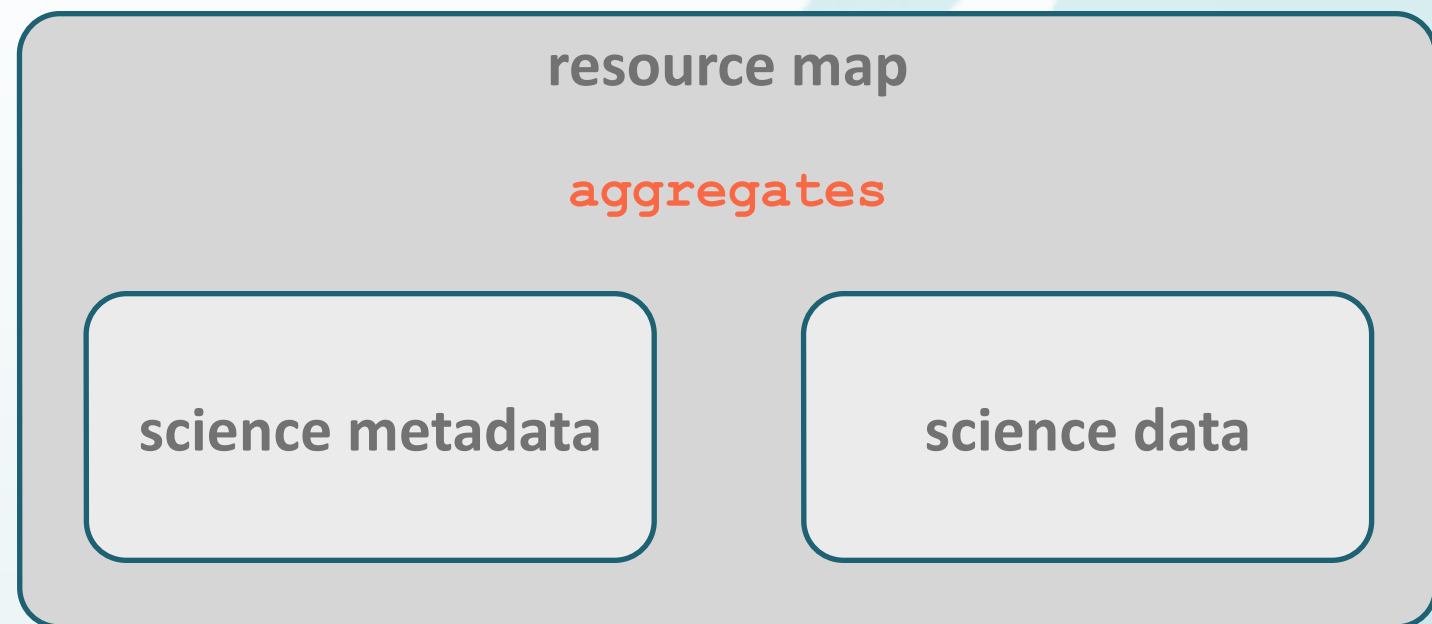
Types of Metadata

Resource Maps



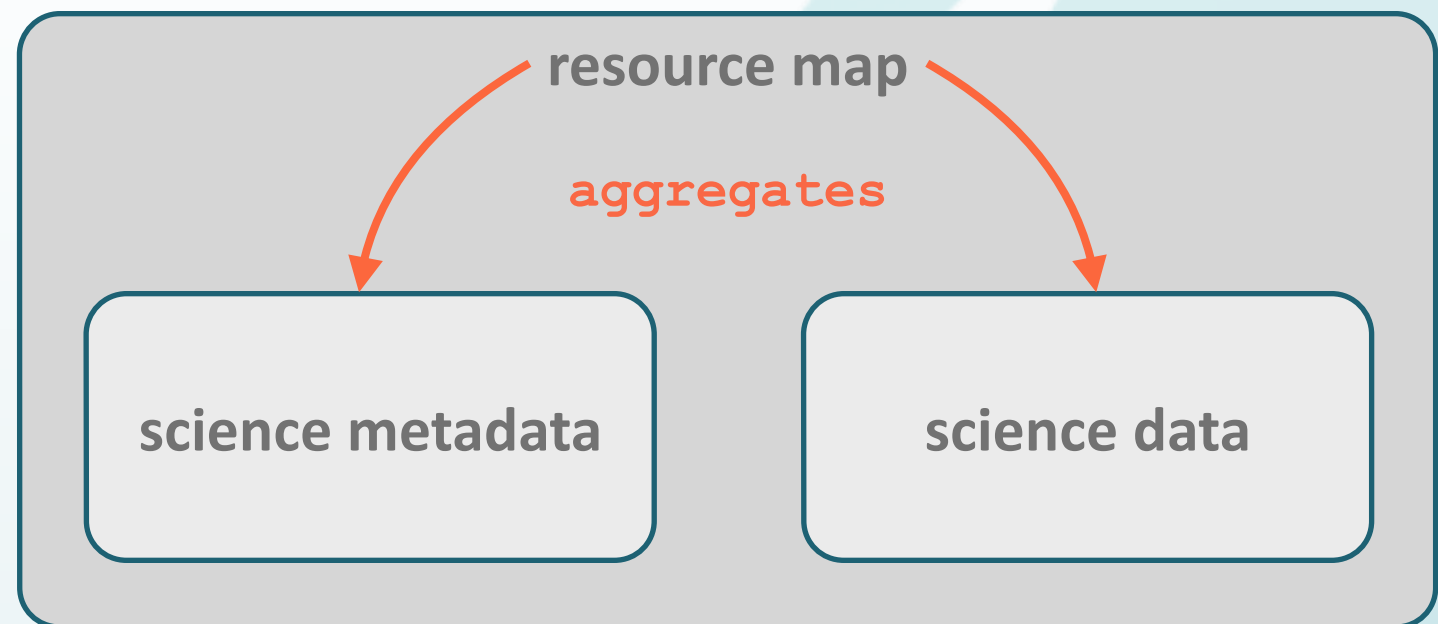
Types of Metadata

Resource Maps



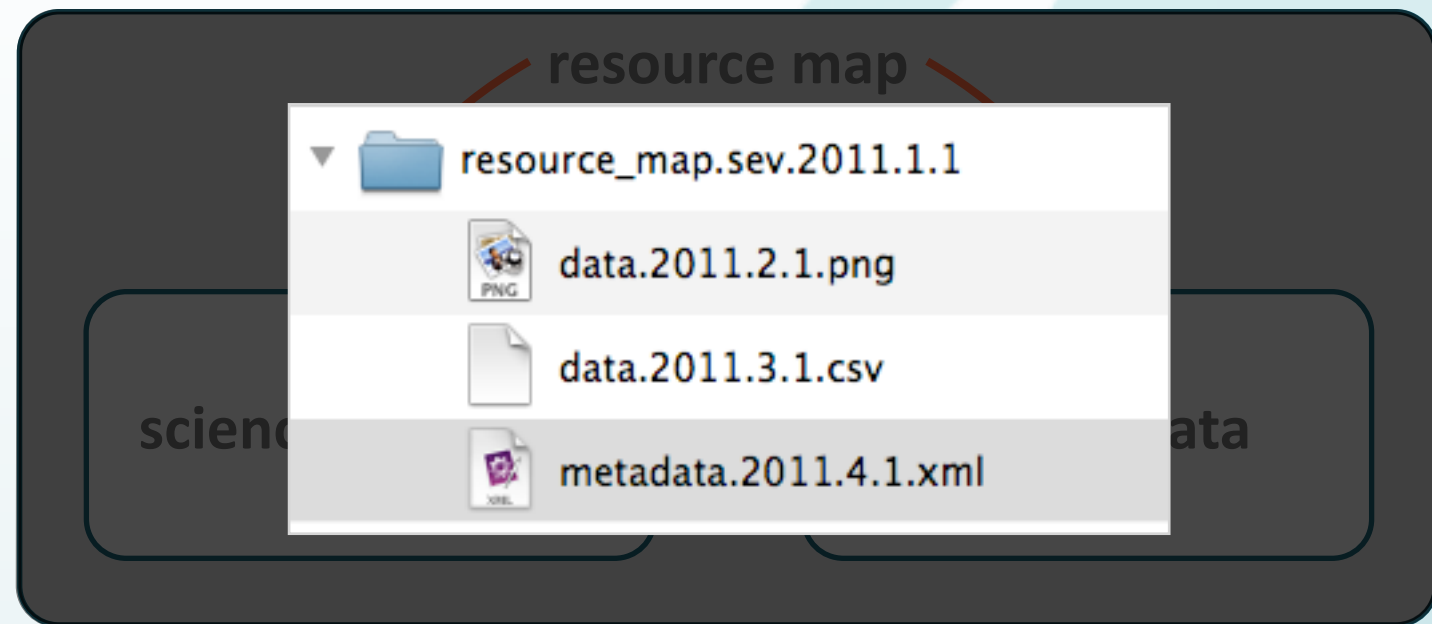
Types of Metadata

Resource Maps



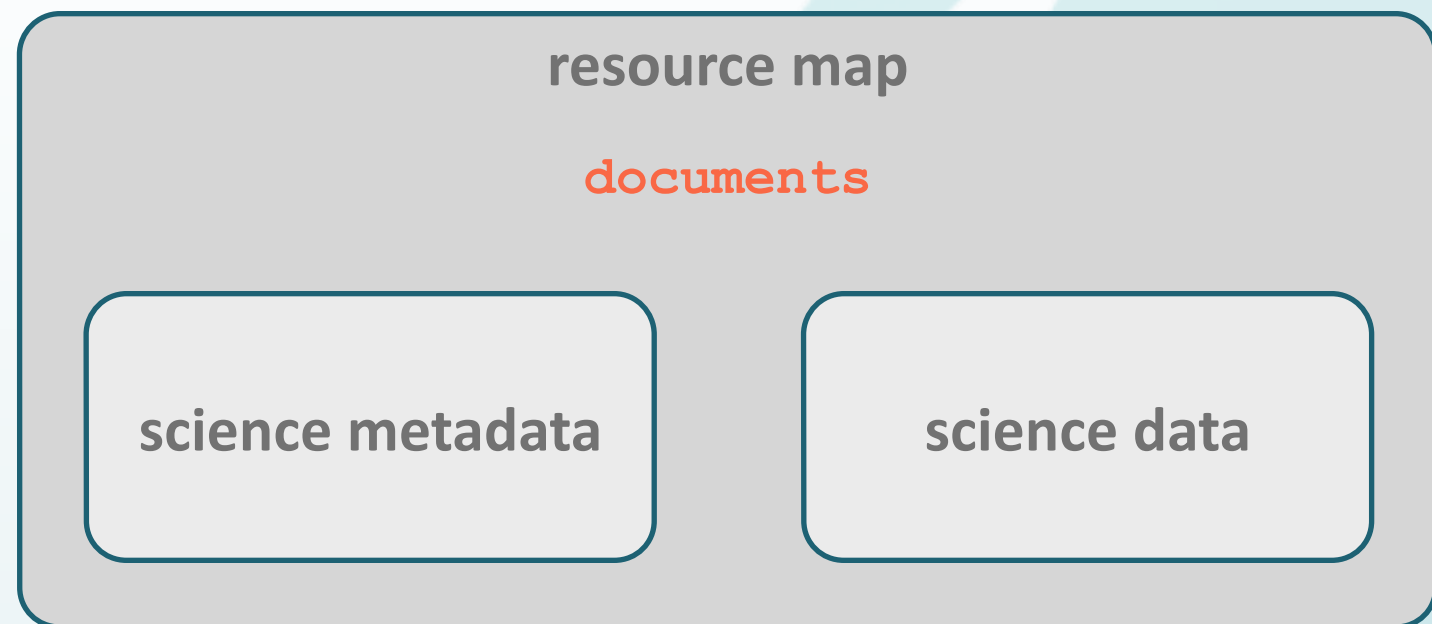
Types of Metadata

Resource Maps



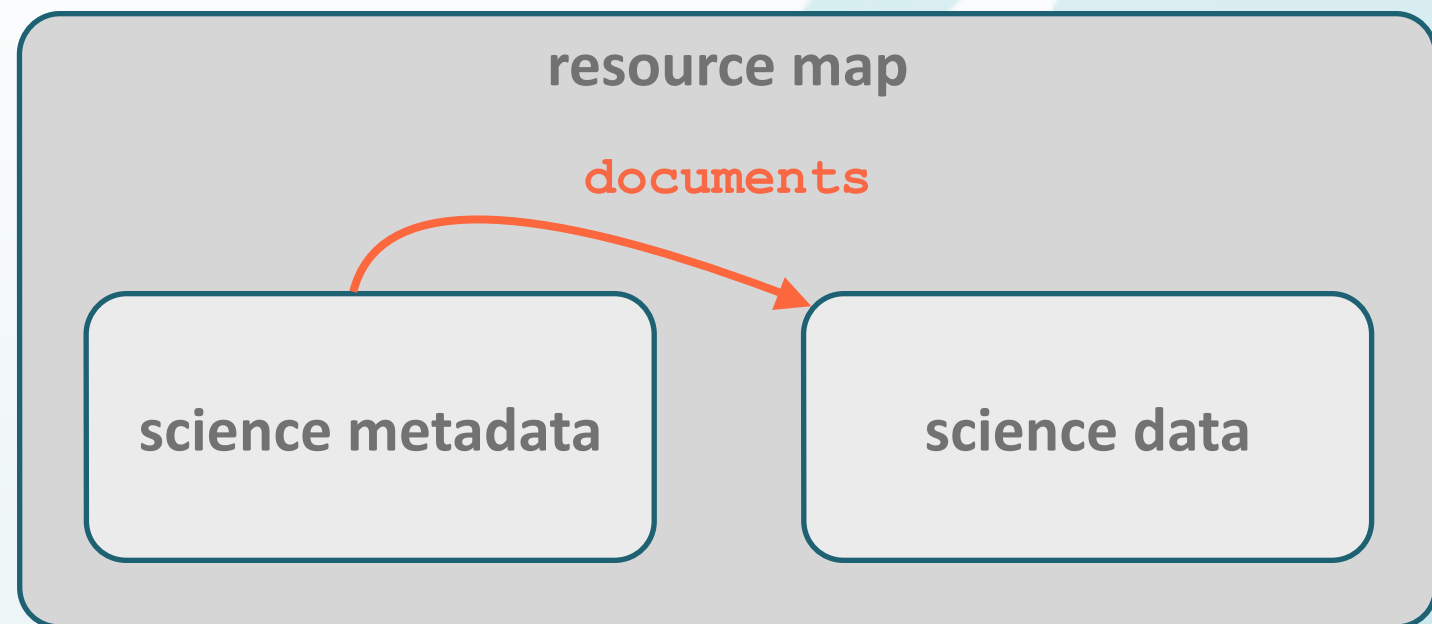
Types of Metadata

Resource Maps



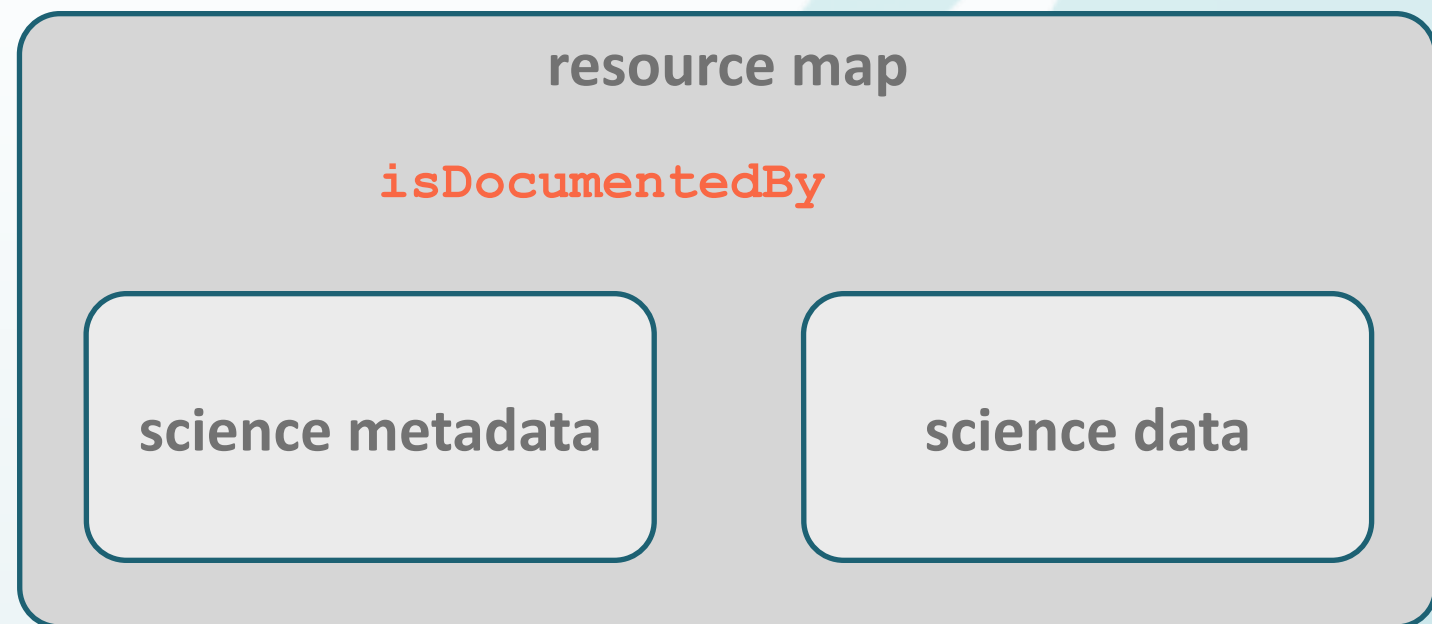
Types of Metadata

Resource Maps



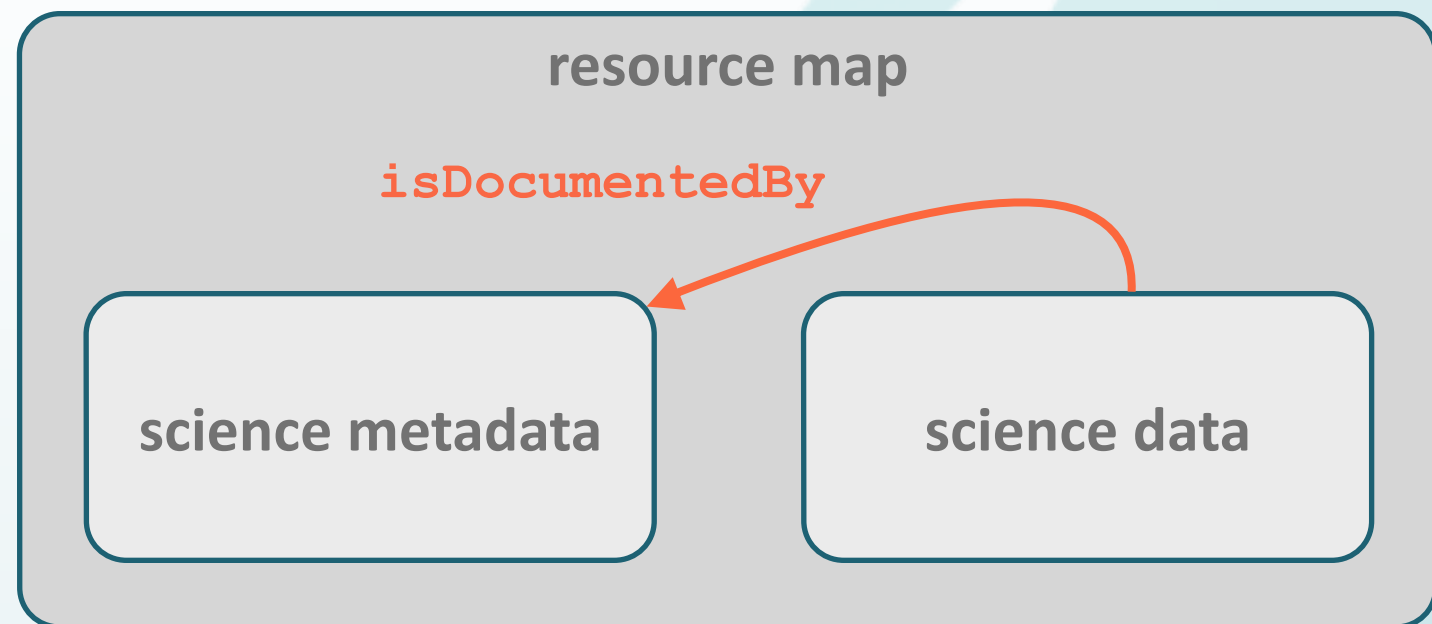
Types of Metadata

Resource Maps



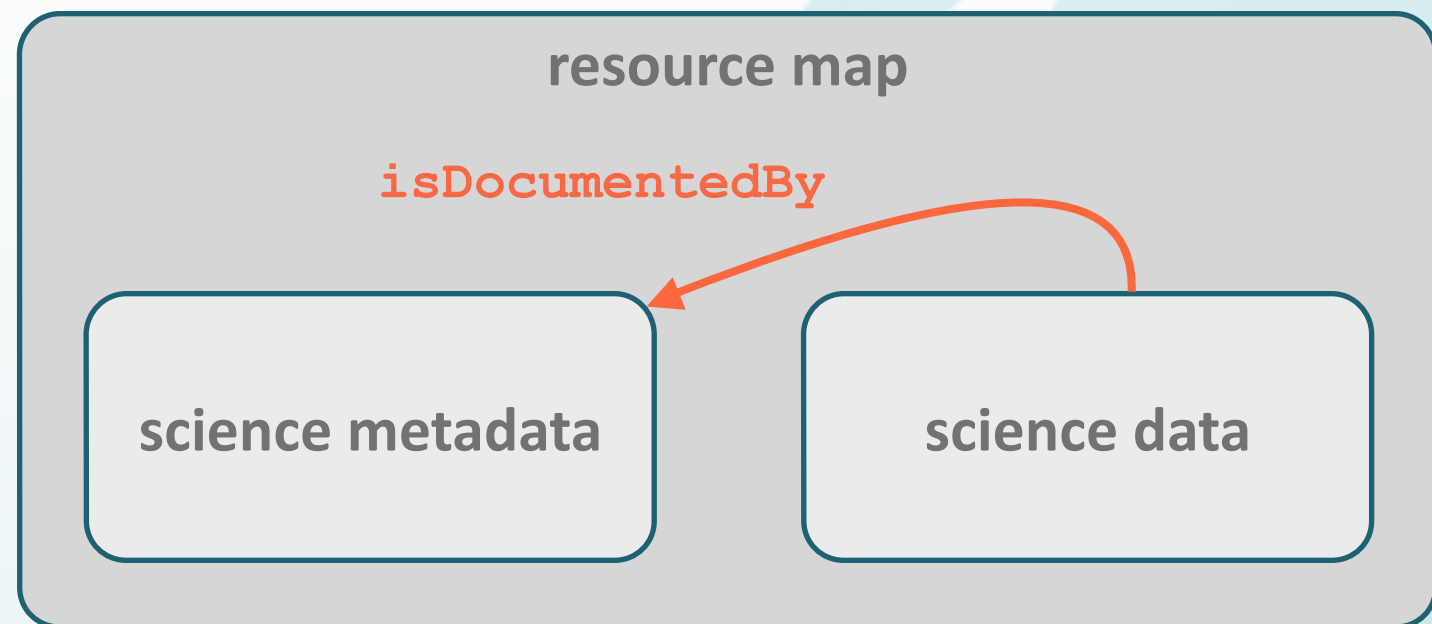
Types of Metadata

Resource Maps



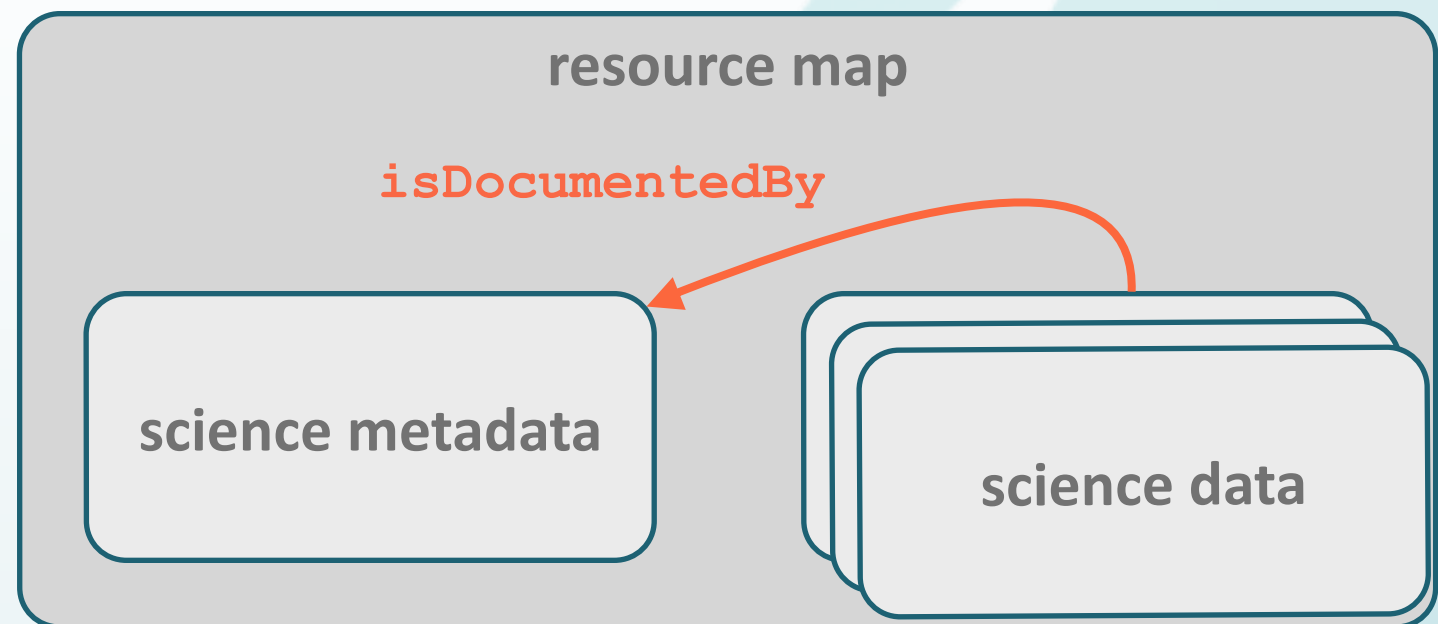
Types of Metadata

Resource Maps



Types of Metadata

Resource Maps



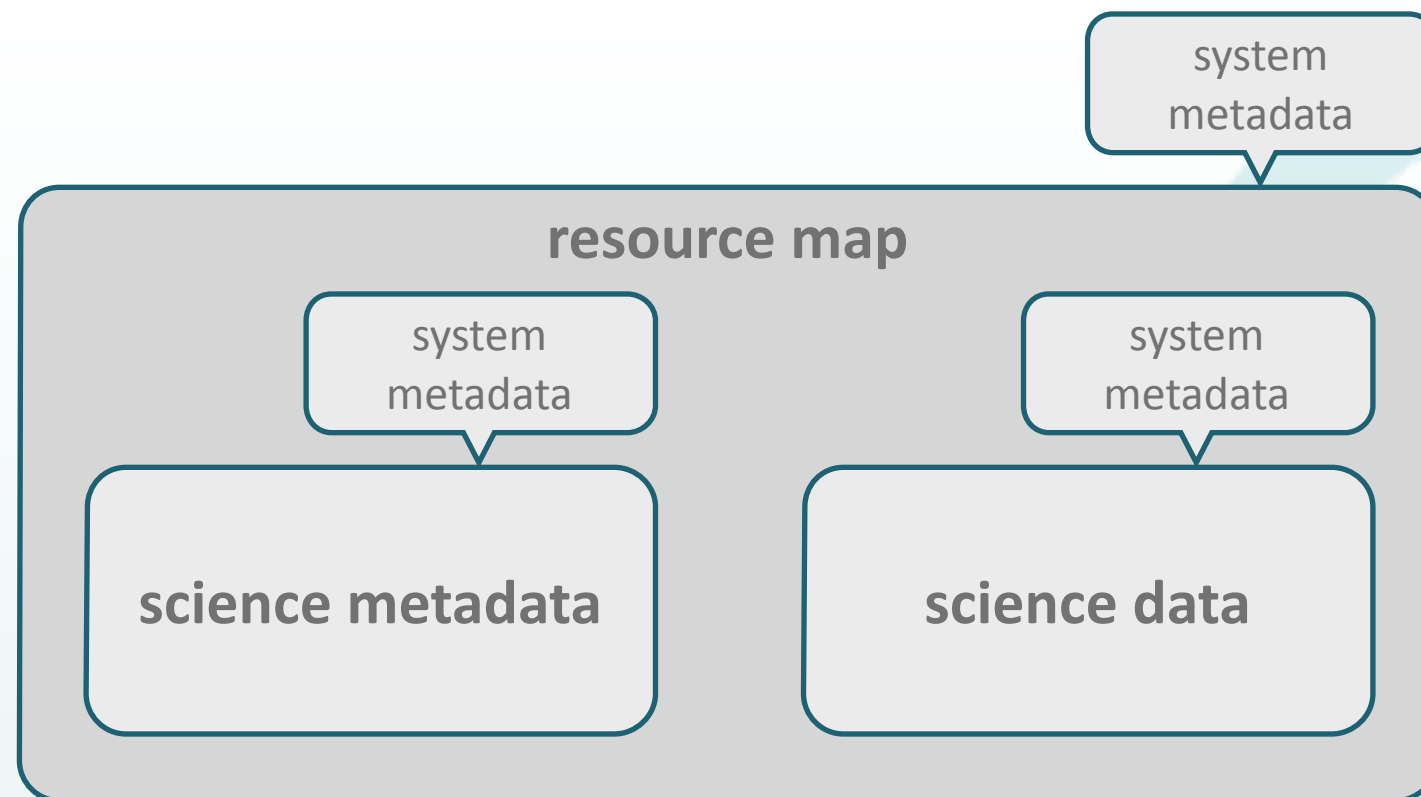
Packaging

Data Package



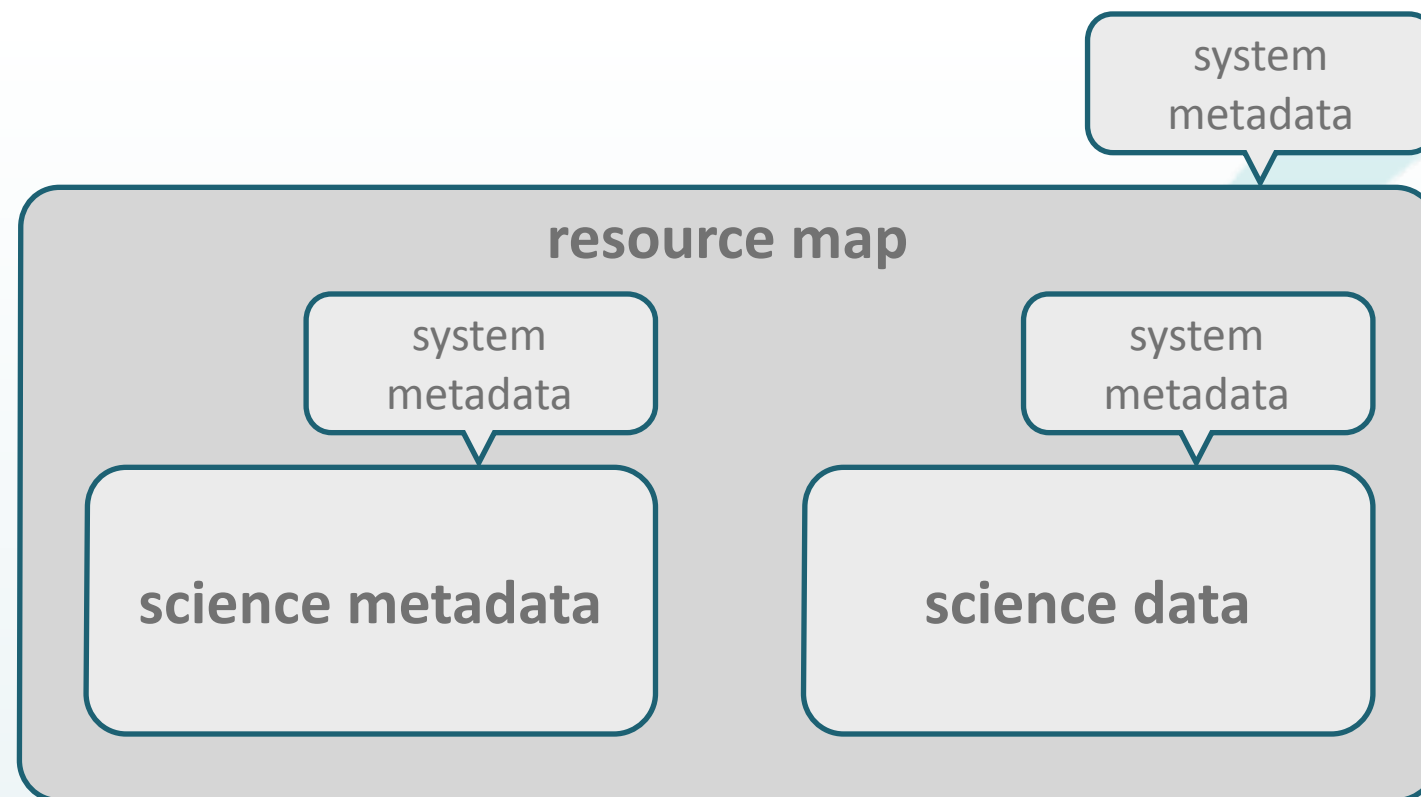
Packaging

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

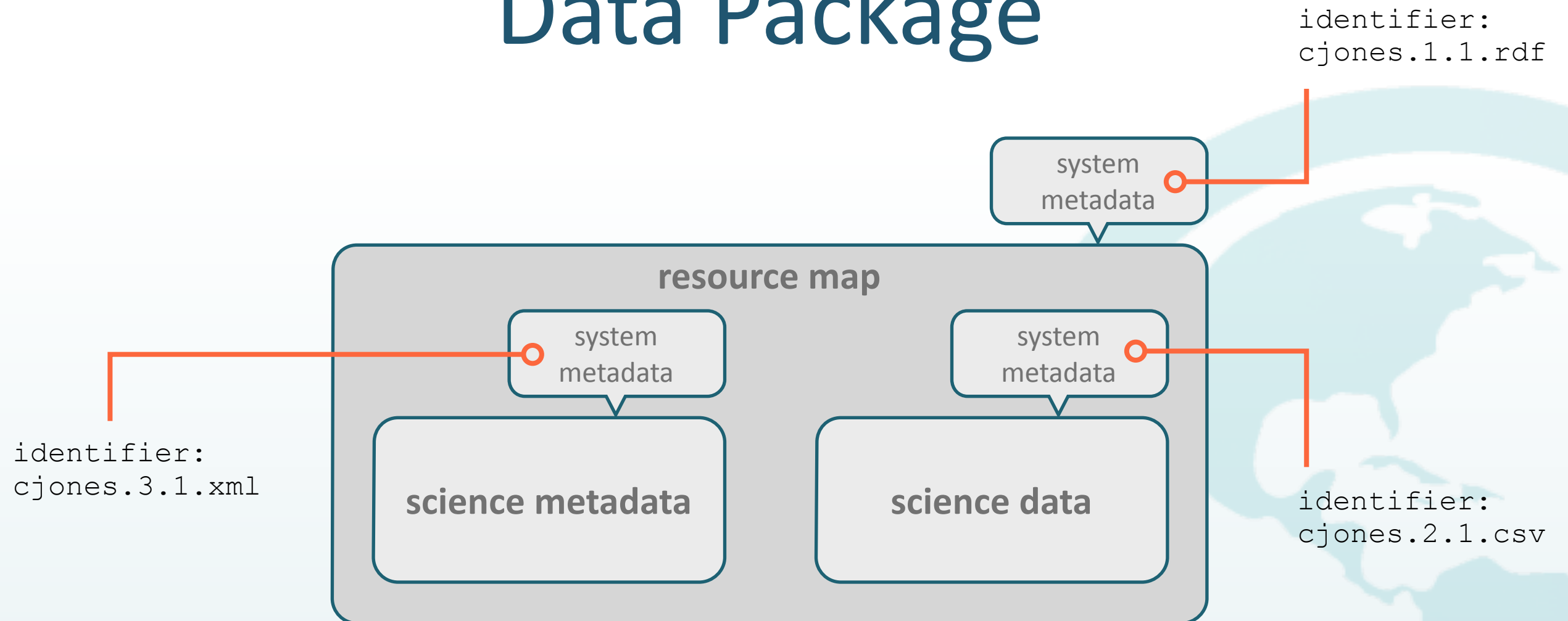
identifier:
cjones.3.1.xml

identifier:
cjones.1.1.rdf

identifier:
cjones.2.1.csv

Identifiers

Data Package



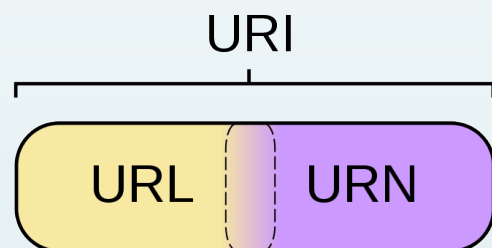
Identifiers



Identifiers

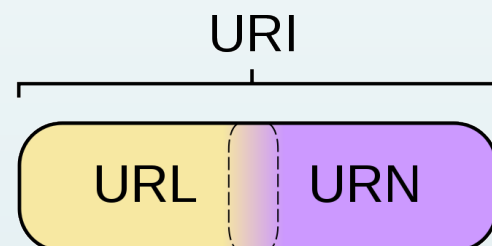
Limited to 800 characters,
no whitespace

Identifiers



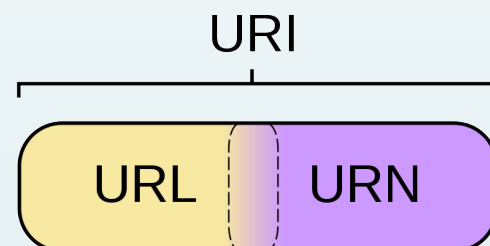
Identifiers

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

doi:10.6073/AA/lake-mendota.2013.1

http://dx.doi.org/10.6073/AA/lake-mendota.2013.1

门多塔.2013.1

68AF6874-6548-4DC7-B798-81B41BB97851

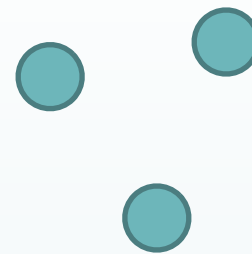
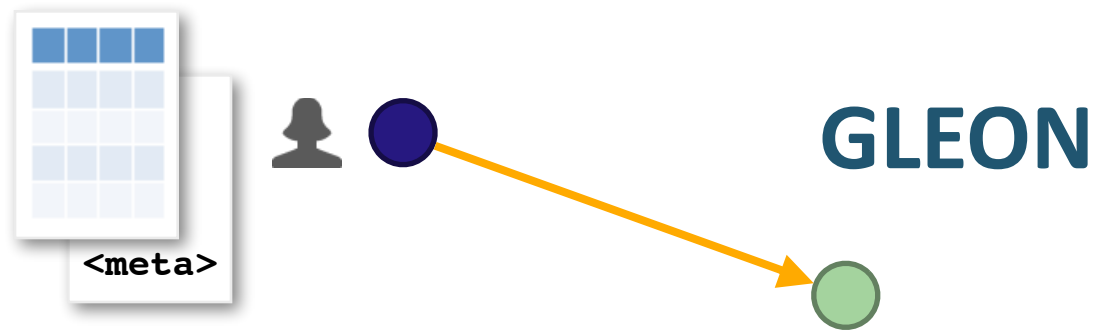
Identifiers



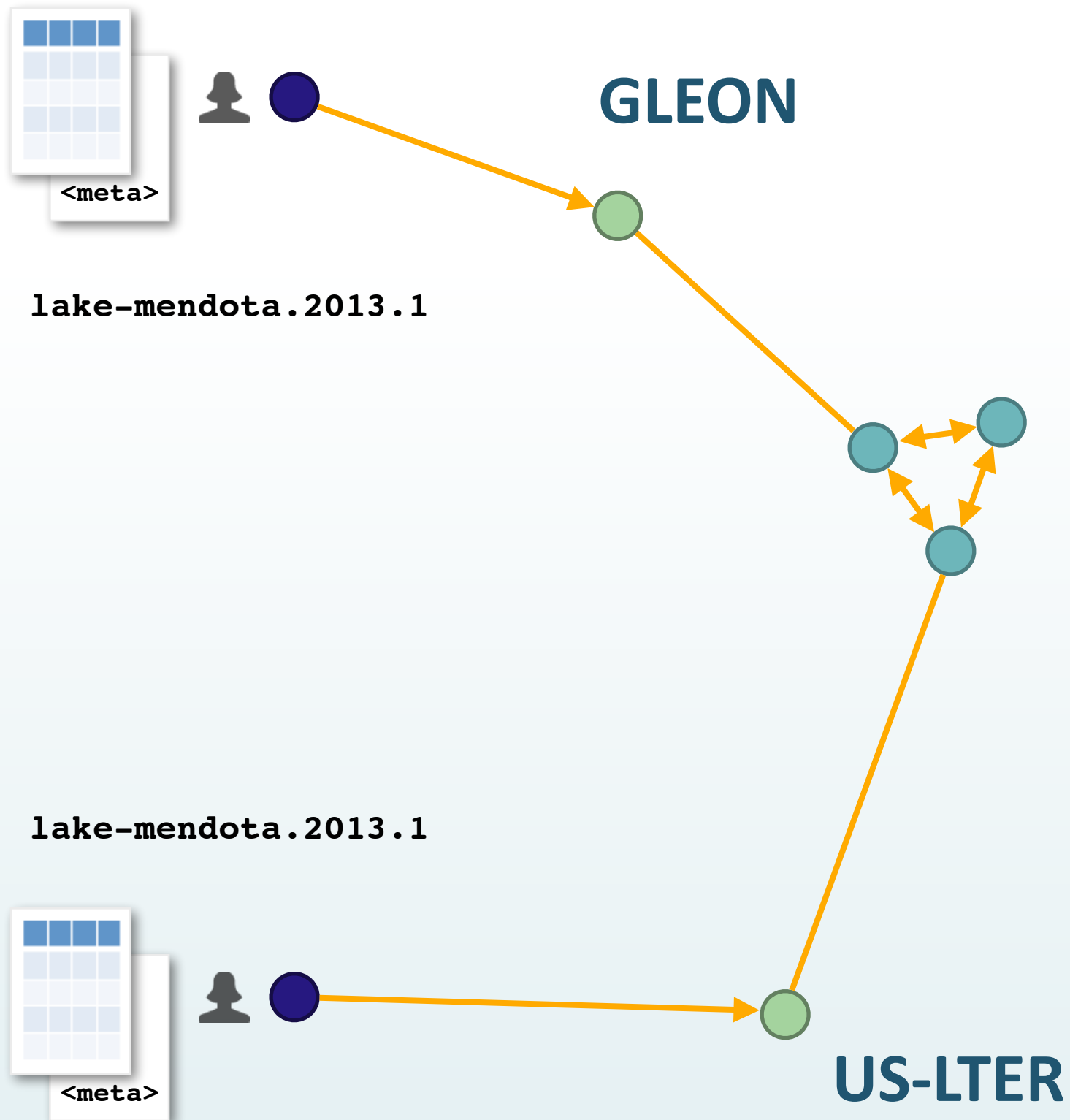
Identifiers

The underlying bytes of an object
referenced by a given identifier
must not change

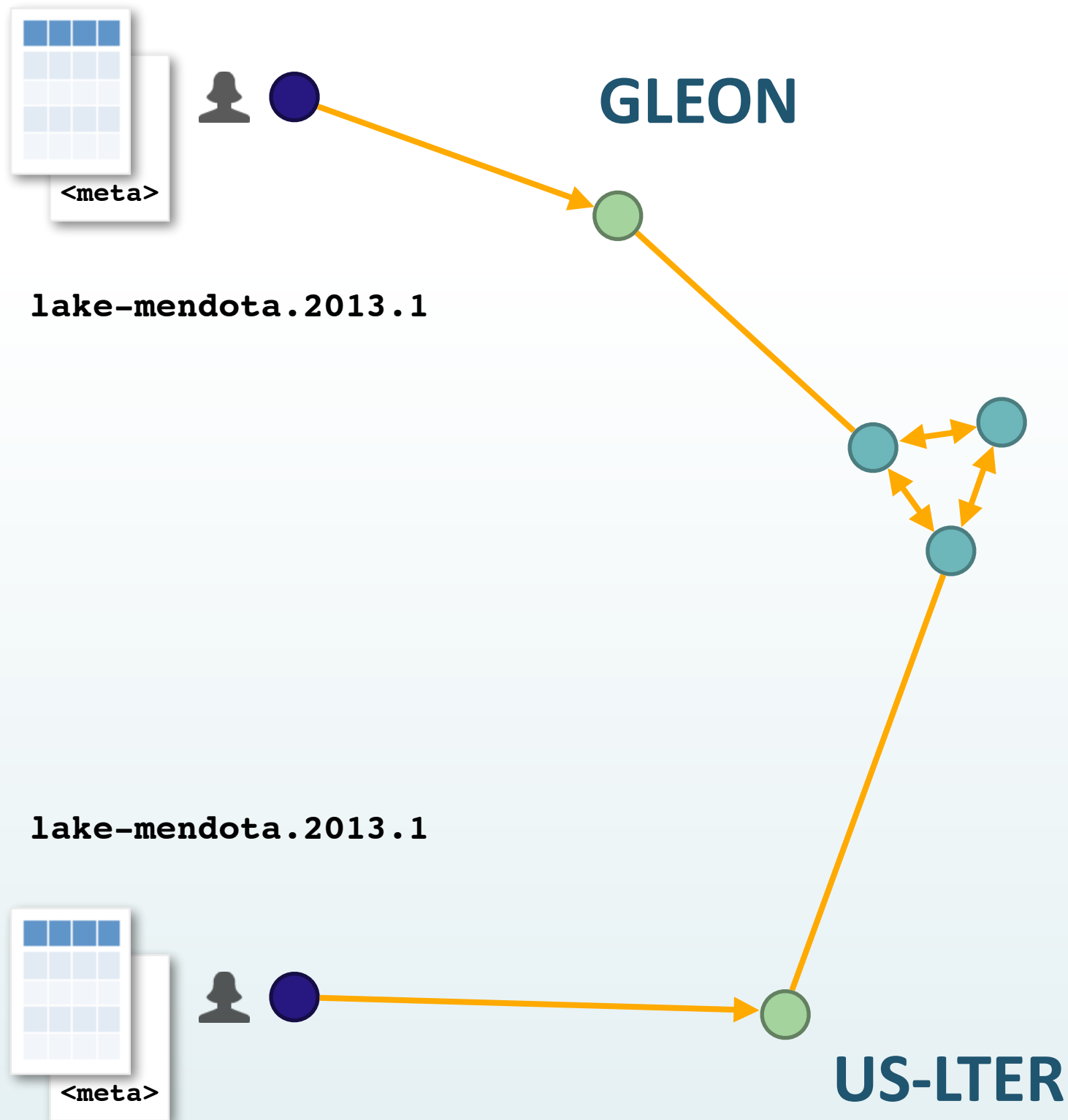
Identifiers



Identifiers



Identifiers



★ 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