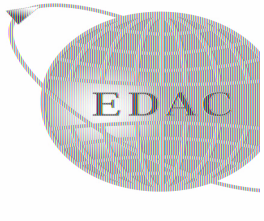
EDAC's Implementation

Soren Scott, Karl Benedict
University of New Mexico



General Requirements

- Data provided through the GSToRE data platform
  - raster data for modeling available through WCS
  - support for multiple documentation standards
- Expanded API to include PROV access
- Develop some method for generating and storing metadata related to PROV and the desired RDF output

Implementing PROV

- Can we reuse existing metadata records?
- Can we leverage working knowledge of metadata practices?
- Can we support a variety of PROV “standards”?
- Can we reduce the overhead for implementing this system?
  - Infrastructure needs?
  - Metadata generation?

Implementing PROV

Can we reuse existing metadata records?

Not entirely. Most of the descriptive elements map to the OWL; however, the workflow structure tends to be incomplete or missing from most metadata records. Or, the workflow structure does not match the ontology so additional work is required.

Implementing PROV

Can we leverage working knowledge of metadata practices?

Yes. The ISO 19115 structure and element concepts are unchanged. But, this depends on the ontology being used. Other standards may not fully support the workflow structure.

No. The workflows reflect issues with collection, education and general documentation practices.

Implementing PROV

Can we support a variety of PROV "standards"?

Yes. Multiple ontologies can be supported either through the generation of ontology-specific DS records or by creating an additional set of wrappers to map the ontology used in the base DS record to the new ontology.

Implementing PROV

Can we reduce the overhead for implementing this system?

Yes. This is a very basic implementation, technically. Once the DS record is generated, PROV (as RDF xml) is generated with a single XSLT transformation. It does not require support for any specific database and can be implemented in any language.

Implementing PROV: Why ISO DS?

We can create one record to document the final data product (our WCS dataset) with internal references, as Data Series MD elements, to more fully capture the workflow and with a ServiceIdentification element to refer to the WCS GetCapabilities request.

- Still have valid ISO
- Still have readable metadata
- Have more complete metadata with well documented intermediate data products

ISO 19115 to PROV

1:1 mappings

Spatial References	ObjectProperty=hasGeospatialProjection Class=Projection DatatypeProperty=hasEPSGCode
Band Identifiers	Class=BandIdentification ObjectProperty=hasBandIdentification ObjectProperty=hasDataBand DatatypeProperty=hasBandName
Units	Class=Units ObjectProperty=hasUnits
Temporal Extents	ObjectProperty=coversTimePeriod ObjectProperty=hasStartDate ObjectProperty=hasEndDate
Spatial Extents	Class=BoxedGeographicRegion objectProperty=coversRegion DatatypeProperty=hasLeftLongitude DatatypeProperty=hasLowerLatitude DatatypeProperty=hasRightLongitude DatatypeProperty=hasUpperLatitude
Data Format	ObjectProperty=encodedInFormat

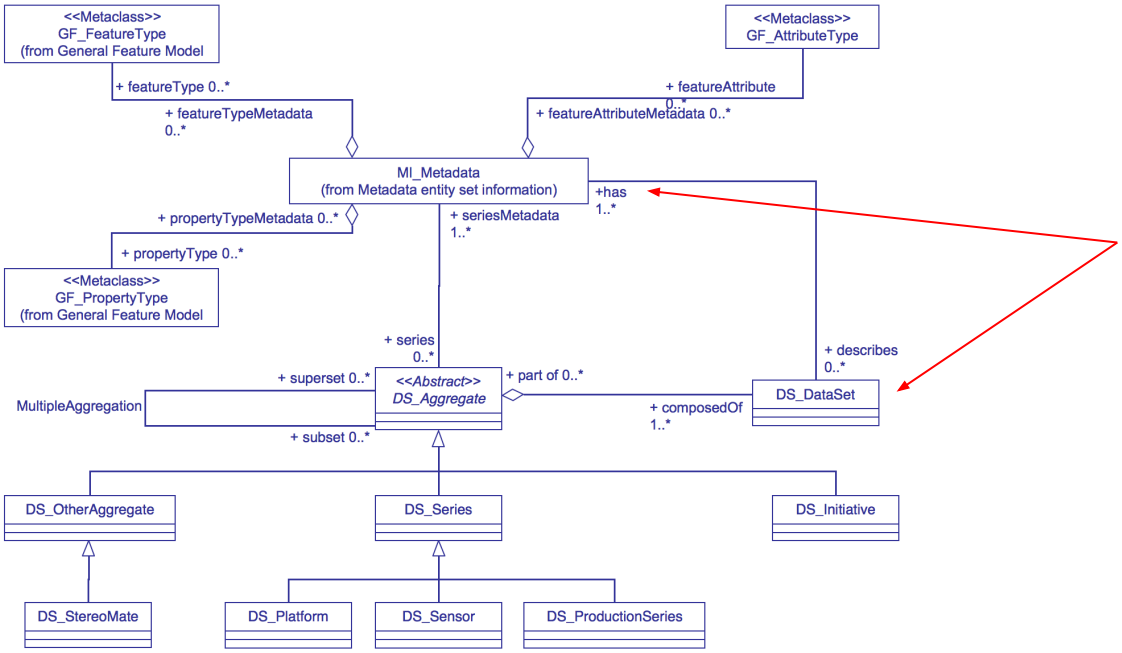
ISO 19115 to PROV

Modified mappings

Identifiers	varies by Owl object and transformation needs
Parameters	keyword element identified by thesaurus (CF or OBOE, etc)
Lineage	main driver for the PROV activities; identify software used, actions taken, input and output sources
Distribution Information	identify the GSTORE service description request
Service Information	identify the WCS as the final data object
Citations	identify the originating source of the data (LP DAAC) and the URL to the data source

Key components - the lineage structure and the MI/MD elements. The specific elements referenced in the MI/MD are related to description and discoverability. And more specifically to the ontologies being used.

ISO 19115 Data Series



Metadata Application

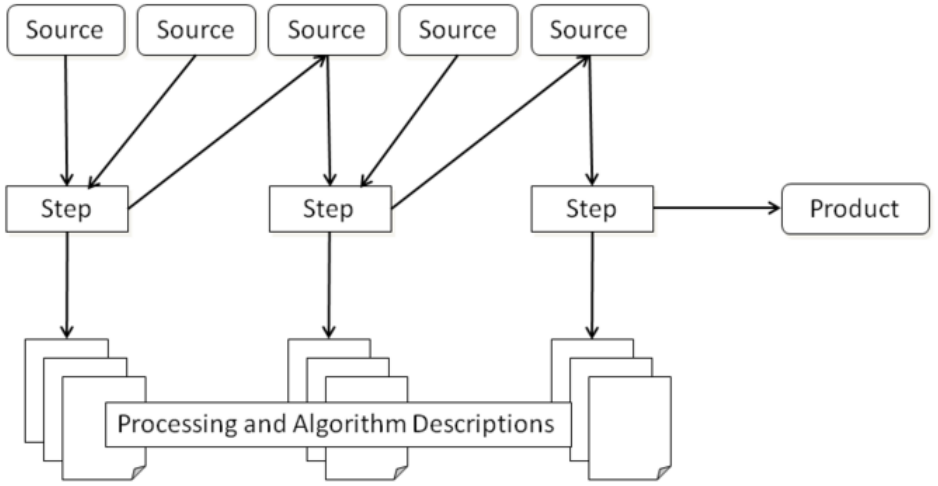
For a published WCS entity represented by the MI record, capture the workflow products from the antecedent source dataset(s) through each process required to publish the WCS.The products are documented by the DS_Dataset MD elements.

ISO 19115 Lineage

Source inputs and outputs are all contained within the lineage, either as a chunk of Sources that are referred to in a processing step or within a processing step.

But the source elements are limited and couldn't capture all of the information we wanted to include in our ontologies.

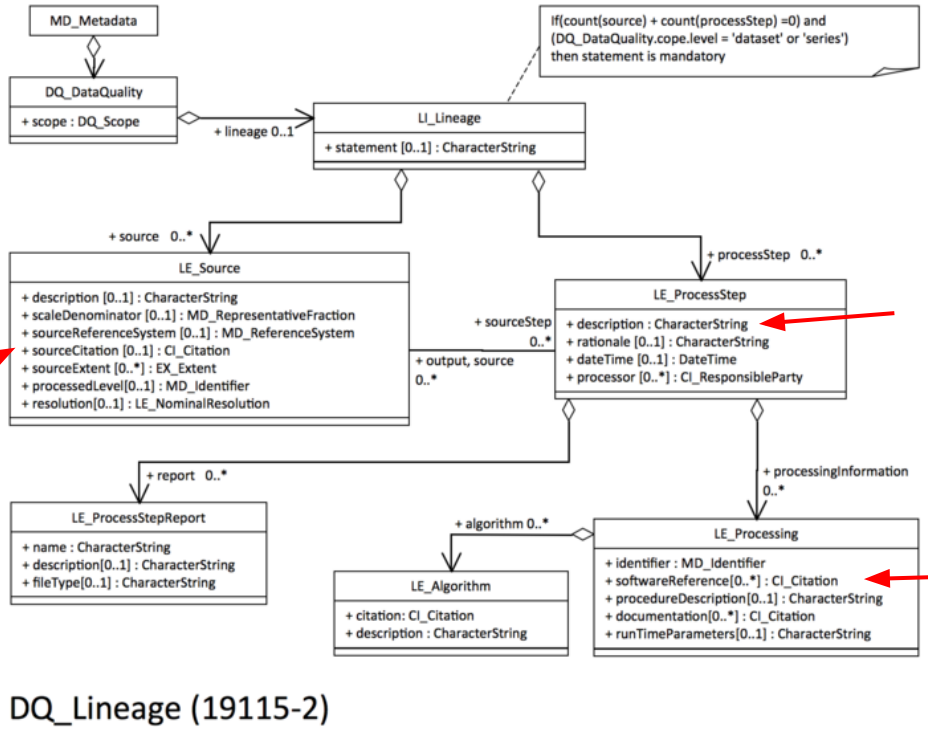
The ISO Lineage Model



via NOAA EDM

ISO 19115 Lineage

LI Lineage



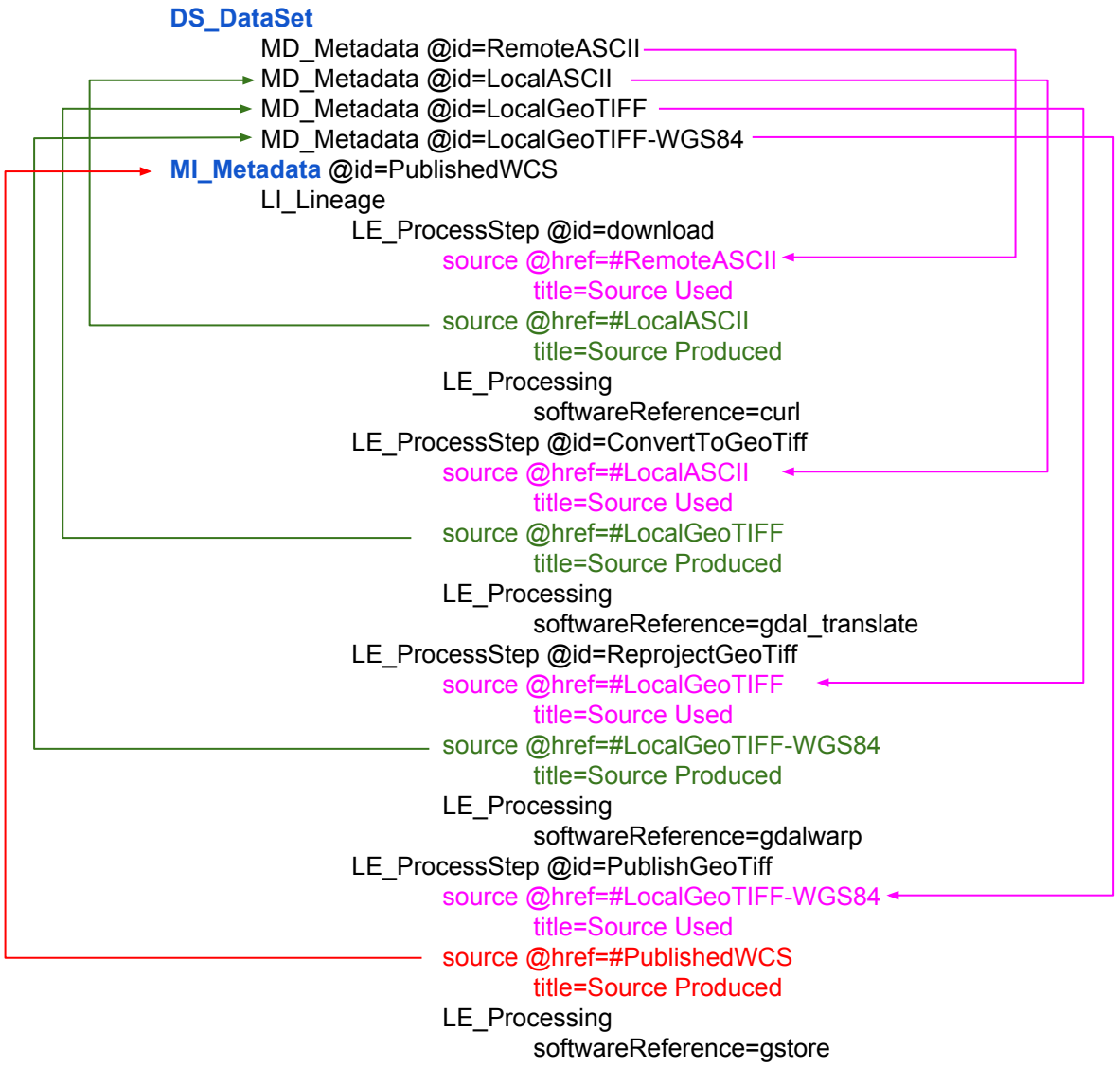
DQ_Lineage (19115-2)

Identify the reference to an MD element in the Data Series; identify whether the source is an input or output; identify the name of the Class

Identify the name of the Activity (using LE_ProcessingStep instead of LI because gmi: processingInformation is invalid in the LI element)

Identify the name of the Software Used

Lineage

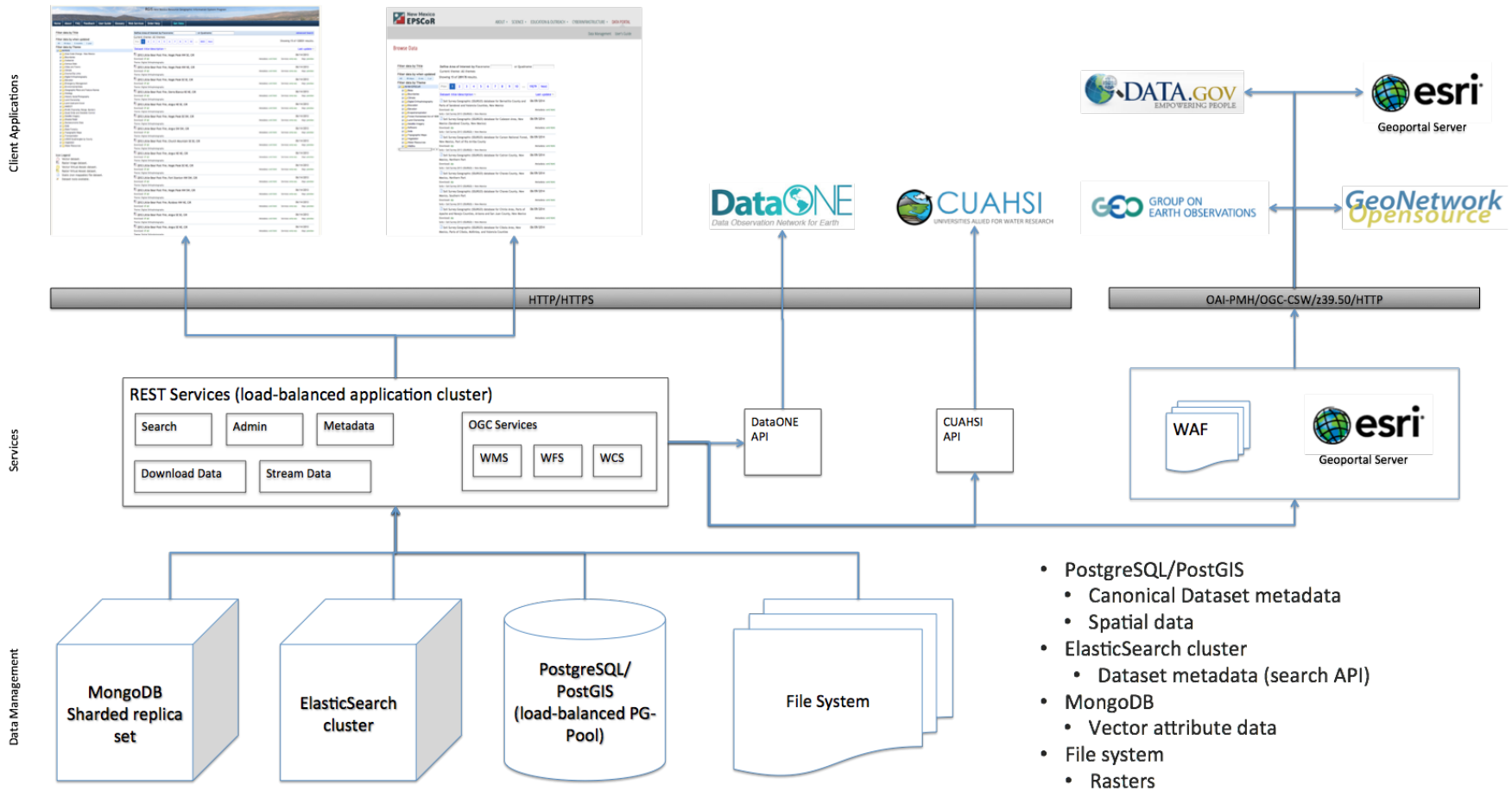

DS_DataSet
MD_Metadata @id=RemoteASCII
MD_Metadata @id=LocalASCII
MD_Metadata @id=LocalGeoTIFF
MD_Metadata @id=LocalGeoTIFF-WGS84
MI_Metadata @id=PublishedWCS
LI_Lineage
LE_ProcessStep @id=download
source @href=#RemoteASCII
title=Source Used
source @href=#LocalASCII
title=Source Produced
LE_Processing
softwareReference=curl
LE_ProcessStep @id=ConvertToGeoTiff
source @href=#LocalASCII
title=Source Used
source @href=#LocalGeoTIFF
title=Source Produced
LE_Processing
softwareReference=gdal_translate
LE_ProcessStep @id=ReprojectGeoTiff
source @href=#LocalGeoTIFF
title=Source Used
source @href=#LocalGeoTIFF-WGS84
title=Source Produced
LE_Processing
softwareReference=gdalwarp
LE_ProcessStep @id=PublishGeoTiff
source @href=#LocalGeoTIFF-WGS84
title=Source Used
source @href=#PublishedWCS
title=Source Produced
LE_Processing
softwareReference=gstore

Generating PROV

At this point, we have a DS record built with the elements required by the ontology and we have a corresponding XSLT for that ontology. PROV is just a simple saxonb-xslt (or XSLT engine of your choice) call.

Data Access

Overview of the GSToRE platform



- PostgreSQL/PostGIS
  - Canonical Dataset metadata
  - Spatial data
- ElasticSearch cluster
  - Dataset metadata (search API)
- MongoDB
  - Vector attribute data
- File system
  - Rasters
  - Compressed datasets
  - Dataset cache

Data Access

1. Search GSToRE for ELSEWeb data:

http://gstore.unm.edu/apps/elseweb/search/datasets.json?version=3

if the dataset supports PROV

if the dataset supports WxS


```
{
    "downloads": [{"tif": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/MYD13Q1_2012009_EVI_MOSAIC.original.tif"}],
    "lastupdate": "20130419",
    "prov": {"modis": {"rdf": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/prov/modis.rdf"}},
    "uuid": "0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3",
    "taxonomy": "geoimage",
    "description": "MODIS Aqua 250m 16 days EVI (2012-01-09 - 2012-01-24)",
    "valid_dates": {
        "start": "20120109",
        "end": "20120124"
    },
    "metadata": [
        {"FGDC-STD-012-2002": {
            "xml": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/metadata/FGDC-STD-012-2002.xml",
            "html": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/metadata/FGDC-STD-012-2002.html"
        }},
        {"ISO-19115:2003": {
            "xml": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/metadata/ISO-19115:2003.xml",
            "html": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/metadata/ISO-19115:2003.html"
        }},
        {
            "ISO-19119:WMS": {"xml": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/metadata/ISO-19119:WMS.xml"},
            "ISO-19119:WCS": {"xml": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/metadata/ISO-19119:WCS.xml"}
        }
    ],
    "metadata-modified": {"all": "2013-04-19T23:40:52Z"},
    "spatial": {
        "epsg": 4326,
        "bbox": [
            -155.5723827,
            20.0001124,
            -74.4920948,
            50
        ]
    },
    "services": [
        {"wms": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/services/ogc/wms?SERVICE=wms&REQUEST=GetCapabilities&VERSION=1.1.1"},
        {"wcs": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/services/ogc/wcs?SERVICE=wcs&REQUEST=GetCapabilities&VERSION=1.1.2"}
    ],
    "preview": "http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/mapper",
    "type": "dataset",
    "id": 348878,
    "categories": [{
        "subtheme": "MODIS EVI",
        "theme": "Vegetation",
        "groupname": "United States"
    }],
    "name": "MYD13Q1_2012009_EVI_MOSAIC"
}
```

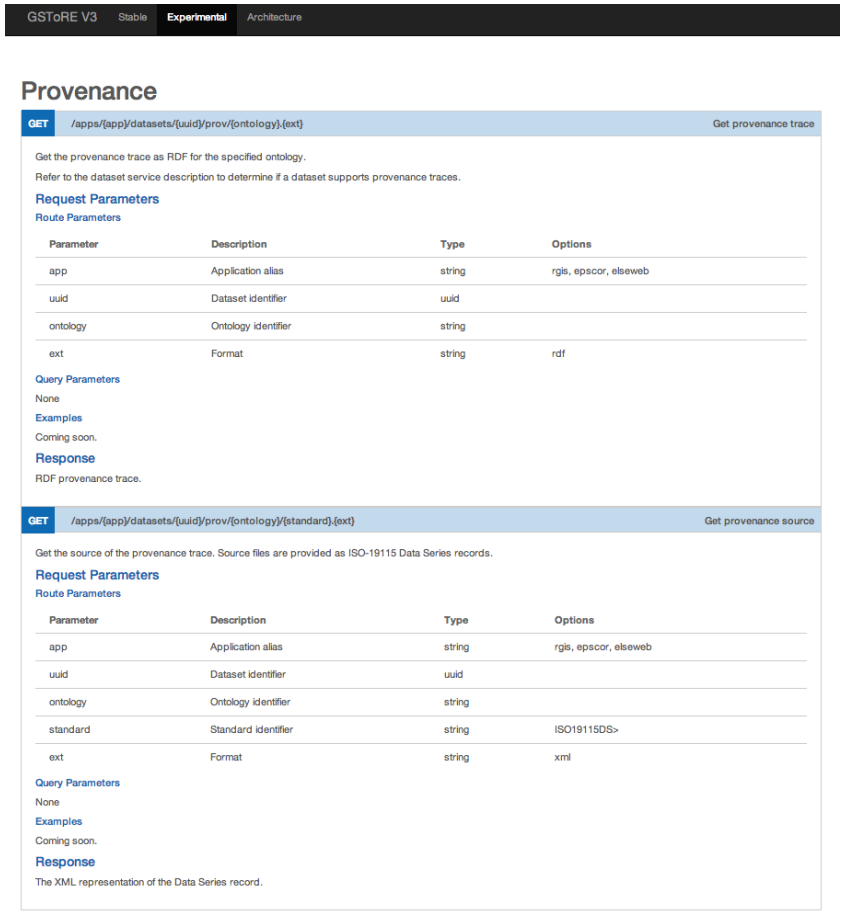
Data Access

Incorporate the metadata link in the WxS GetCapabilities response using

- wxs_metadatalink_href (URL)
- wxs_metadatalink_format (mimetype)
- wxs_metadatalink_type (standard)

```
<ows:Metadata xlink:type="simple" xlink:href="http://gstore.unm.edu/apps/elseweb/datasets/0caa249d-bf8a-4f53-bdfb-0b12ca4cf2f3/metadata/FGDC-STD-012-2002.xml" about="FGDC-STD-012-2002" xlink:role="text/xml"/>
```

Data Access



GSToRE V3 Stable **Experimental** Architecture

Provenance

GET /apps/{app}/datasets/{uuid}/prov/{ontology}.{ext} — Get provenance trace

Get the provenance trace as RDF for the specified ontology.

Refer to the dataset service description to determine if a dataset supports provenance traces.

Request Parameters

Route Parameters

Parameter	Description	Type	Options
app	Application alias	string	rgis, epscor, elseweb
uuid	Dataset identifier	uuid	
ontology	Ontology identifier	string	
ext	Format	string	rdf

Query Parameters

None

Examples

Coming soon.

Response

RDF provenance trace.

GET /apps/{app}/datasets/{uuid}/prov/{ontology}/{standard}.{ext} — Get provenance source

Get the source of the provenance trace. Source files are provided as ISO-19115 Data Series records.

Request Parameters

Route Parameters

Parameter	Description	Type	Options
app	Application alias	string	rgis, epscor, elseweb
uuid	Dataset identifier	uuid	
ontology	Ontology identifier	string	
standard	Standard identifier	string	ISO19115DS>
ext	Format	string	xml

Query Parameters

None

Examples

Coming soon.

Response

The XML representation of the Data Series record.

Modifications to the core GSTORE platform:

- two access routes: one for the base ISO Data Series metadata and one for the PROV rdf generated from the DS
- internal schema changes to handle multiple ontologies per dataset
- additional metadata cache for the DS records

Examples

ISO DS records and PROV (rdf):

http://roomthily.github.io/gstore-elseweb-preview

Transformations:

http://github.com/roomthily/gstore-elseweb

Limitations

- Must support ISO 19115 DS
- Most existing metadata doesn't fully implement the expected lineage structure
- ISO DS is specific to the ontology - certain element values must match something in the associated OWL
  - requires multiple DS records to support multiple ontogies/semantic web platforms
- The OWL isn't identified in the ISO DS
- Does not currently support access to the intermediate data objects or Activity/Entity objects.
  - the API doesn't provide access to a specific instance of a SoftwareEntity used by a specific Mosaicking activity for a specific WCS, for example

Limitations

- Can’t validate the ISO DS against the structure inherent in the ontology
- Can’t validate the element values against the definitions in the OWL
  - CodeLists have some potential but are not available for most elements
- Can’t validate the generated RDF