

ISO Lineage

ISO Lineage

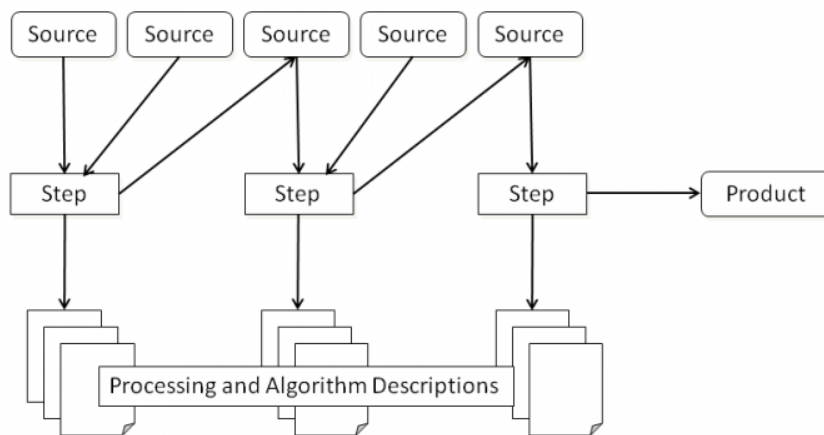
Tracking data sources and processing done to them is becoming increasingly important as scientists seek to define trends and unexpected changes in the environment. Keeping track of data transformations and processing, generally termed *lineage*, is an important role of high-quality metadata. The ISO metadata standard provides a simple lineage model based on *sources* which are either used or produced in a series of *process steps*. This model can be helpful in many cases despite its simplicity. Sources and process steps are linked together to describe the lineage of a resource.

FGDC and ISO source descriptions have several important differences.

1. The FGDC source model is based on citations to scientific papers. These citations do not include contact information (commonly an email address or a URL) which limits their usefulness in many cases.
2. The scope of the references to the citations is limited to the metadata lineage section. This makes it difficult to unambiguously identify sources across multiple records.
3. The FGDC standard allows the description of the temporal extent of a source, but not the spatial extent.

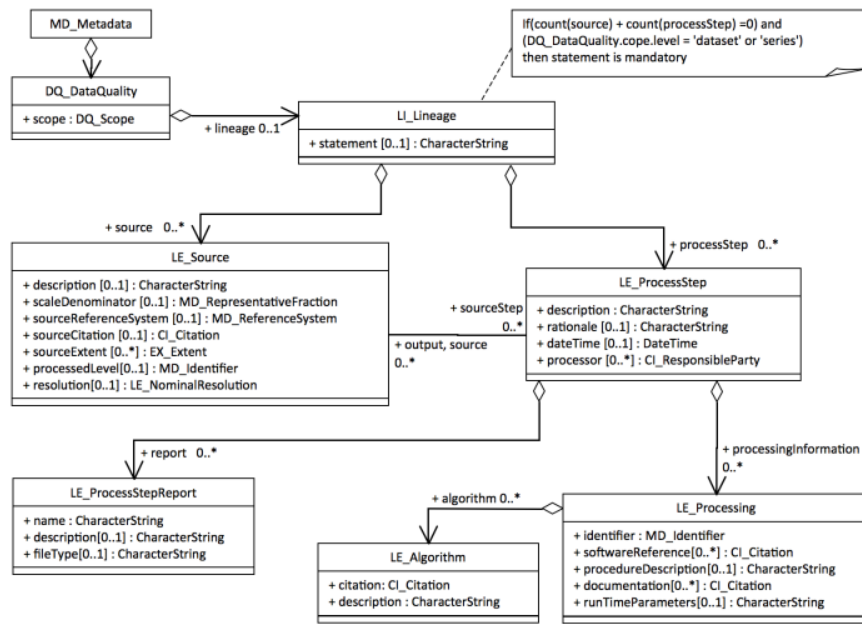
All of these limitations are overcome in the ISO lineage model.

The Model



ISO Lineage Model

This Figure shows an overview of the ISO lineage model which links sources to process steps. Sources can be input to, or output (19115-2) from a process step. Each process step has associated processing and algorithm information (also added in 19115-2). These improvements make it important to use 19115-2 if you need good lineage descriptions.



DQ_Lineage (19115-2)

ISO 19115-2 Lineage UML

This Figure shows more detail in the UML model used by the ISO Standard to describe lineage. In some cases, a simple descriptive statement can describe the lineage effectively. In more complex cases, multiple sources and process steps might be required. The definitions of sources and processSteps are also shown in the UML. The capability to specify the spatial and temporal extent of the source and to describe the rationale for a process step are new in the ISO Standard. Note that each source can have any number of associated sourceSteps and that each processStep can have any number of sources (and outputs in ISO 19115-2).

Sources and Steps

The original ISO 19115 Source descriptions (LI_Source) were extended in 19115-2 to include several more elements. The LE_Source includes the following elements:

```

LE_Source
+ description[0..1]: CharacterString
+ scaleDenominator[0..1]: MD_RepresentativeFraction
+ sourceReferenceSystem[0..1]: MD_ReferenceSystem
+ sourceCitation[0..1]: CI_Citation
+ sourceExtent[0..*]: EX_Extent
+ processedLevel[0..1]: MD_Identifier
+ resolution[0..1]: LE_NominalResolution
  
```

and Process Steps include

```

LE_ProcessStep
+ description: CharacterString
+ rationale[0..1]: CharacterString
+ dateTime[0..1]: DateTime
+ processor[0..*]: CI_ResponsibileParty
+ processingInformation[0..*]: LE_Processing
+ report[0..*]: LE_ProcessStepReport
  
```

```

LE_Processing
+ identifier: MD_Identifier
+ softwareReference[0..*]: CI_Citation
+ procedureDescription[0..1]: CharacterString
+ documentation[0..*]: CI_Citation
+ runTimeParameters[0..1]: CharacterString
+ algorithm[0..*]: LE_Algorithm
  
```

```

LE_Algorithm
+ citation: CI_Citation
+ description: CharacterString
  
```

```

LE_ProcessStepReport
+ name: CharacterString
+ description[0..1]: CharacterString
+ fileType[0..1]: CharacterString
  
```

The ISO Lineage model is simple but is probably sufficient for many common processing scenarios. It may only provide summary information in complex processing scenarios. This is facilitated by the use of CI_Citations in LE_Sources, LE_Processing, and LE_Algorithm. The resources referenced by these citations can provide more detail when necessary.

XML Implementation

Implementing these relationships in XML can seem daunting. It is accomplished in the XML representation using ids and references. The LE_Source and LE_ProcessStep objects (the boxes in the UML) are implemented as independent children of the LI_Lineage object with unique identifiers and the relationships, the source, output, and sourceStep roles in between the boxes, are implemented as references.

The example shown below shows the lineage section of a DART metadata record. This DART dataset is made up of data from three different deployments. Each of these is listed as a source in the second part of the lineage section. Each source includes

- an id (D165_1999, D165_2000, and D165_2001),
- a spatial and temporal extent defined using a reference to a full description in a different part of the record ([xlink:href="#Extent_D165_2001"](#)), and
- a sourceStep which is also defined by a reference to a full definition located in the first part of the lineage section (e.g. [xlink:href="#Received_D165_2001"](#)).

The processing of each source is described in the first part of the lineage section. In this case, the process is the receipt of the the data by the archive. The processSteps include:

- a brief description of the process
- when it was done
- who did it, defined by a reference to the seriesmetadataContact defined elsewhere in the record,
- a reference to the source that was processed ([gmd:source xlink:href="#D165_1999"](#)).

Note the use of id's within this record to identify sources and process steps and to make links between them.

```

<gmd:dataQualityInfo>
  <gmd:DQ_DataQuality>
    <gmd:scope>
      <gmd:DQ_Scope id="datasetScope">
        <gmd:level>
          <gmd:MD_ScopeCode codeList="./resources/codeList.xml#MD_ScopeCode" codeListValue="dataset"/>
        </gmd:level>
        <gmd:extent xlink:href="#boundingExtent"/>
      </gmd:DQ_Scope>
    </gmd:scope>
    <gmd:lineage xlink:title="Dart Bouy D165 Processing">
      <gmd:LI_Lineage uuid="95BD4CCC-D27D-8DE4-E040-0AC8C5BB43B64">
        <gmd:statement>
          <gco:CharacterString>Dart Bouy D165 Processing</gco:CharacterString>
        </gmd:statement>
        <gmd:processStep>
          <gmd:LI_ProcessStep id="Received_D165_1999">
            <gmd:description>
              <gco:CharacterString>Received edited data D165_1999-ed</gco:CharacterString>
            </gmd:description>
            <gmd:dateTime>
              <gco:DateTime>2005-09-02T00:00:00</gco:DateTime>
            </gmd:dateTime>
            <gmd:processor xlink:href="#seriesMetadataContact"/>
            <gmd:source xlink:href="#D165_1999"/>
          </gmd:LI_ProcessStep>
        </gmd:processStep>
        <gmd:processStep>
          <gmd:LI_ProcessStep id="Received_D165_2000">
            <gmd:description>
              <gco:CharacterString>Received edited data D165_2000-ed</gco:CharacterString>
            </gmd:description>
            <gmd:dateTime>
              <gco:DateTime>2005-09-02T00:00:00</gco:DateTime>
            </gmd:dateTime>
            <gmd:processor xlink:href="#seriesMetadataContact"/>
            <gmd:source xlink:href="#D165_2000"/>
          </gmd:LI_ProcessStep>
        </gmd:processStep>
        <gmd:processStep>
          <gmd:LI_ProcessStep id="Received_D165_2001">
            <gmd:description>
              <gco:CharacterString>Received edited data D165_2001-ed</gco:CharacterString>
            </gmd:description>
            <gmd:dateTime>
              <gco:DateTime>2005-09-02T00:00:00</gco:DateTime>
            </gmd:dateTime>
            <gmd:processor xlink:href="#seriesMetadataContact"/>
            <gmd:source xlink:href="#D165_2001"/>
          </gmd:LI_ProcessStep>
        </gmd:processStep>
        <gmd:source>
          <gmd:LI_Source id="D165_1999">
            <gmd:description>
              <gco:CharacterString>gov.noaa.ngdc.dart:D165_1999</gco:CharacterString>
            </gmd:description>
            <gmd:sourceExtent xlink:href="#Extent_D165_1999"/>
            <gmd:sourceStep xlink:href="#Received_D165_1999"/>
          </gmd:LI_Source>
        </gmd:source>
        <gmd:source>
          <gmd:LI_Source id="D165_2000">
            <gmd:description>
              <gco:CharacterString>gov.noaa.ngdc.dart:D165_2000</gco:CharacterString>
            </gmd:description>
            <gmd:sourceExtent xlink:href="#Extent_D165_2000"/>
            <gmd:sourceStep xlink:href="#Received_D165_2000"/>
          </gmd:LI_Source>
        </gmd:source>
        <gmd:source>
          <gmd:LI_Source id="D165_2001">
            <gmd:description>
              <gco:CharacterString>gov.noaa.ngdc.dart:D165_2001</gco:CharacterString>
            </gmd:description>
            <gmd:sourceExtent xlink:href="#Extent_D165_2001"/>
            <gmd:sourceStep xlink:href="#Received_D165_2001"/>
          </gmd:LI_Source>
        </gmd:source>
      </gmd:LI_Lineage>
    </gmd:lineage>
  </gmd:DQ_DataQuality>
</gmd:dataQualityInfo>

```

This XML shows parts of a lineage section for a CoastWatch Swath dataset (see [entire record](#)).

```
<gmd:lineage>
  <gmd:LI_Lineage>
    <gmd:processStep>
      <gmd:LI_ProcessStep id="121">
        <gmd:description>
          <gco:CharacterString>
            * Ingest and calibrate: ingests raw satellite data to TeraScan data format.* Automatic navigation:
corrects an ingested AVHRR pass file.
          </gco:CharacterString>
        </gmd:description>
        <gmd:dateTime gco:nilReason="Not complete"/>
        <gmd:processor>...</gmd:processor>
        <gmd:source xlink:href="#HRPT_AVHRR_L0"/> <!-- 19115-2: input -->
        <gmd:source xlink:href="#HRPT_AVHRR_L1B"/> <!-- 19115-2: input -->
        <gmd:source xlink:href="#TDF_Temp"/> <!-- 19115-2: output -->
      </gmd:LI_ProcessStep>
    </gmd:processStep>
    <gmd:processStep>
      <gmd:LI_ProcessStep id="122">
        ...
        <gmd:source xlink:href="#TDF_Temp"/> <!-- 19115-2: input -->
        <gmd:source xlink:href="#SST_Cloud_TDF"/> <!-- 19115-2: output -->
      </gmd:LI_ProcessStep>
    </gmd:processStep>
    <gmd:processStep>...</gmd:processStep>
    <gmd:processStep>...</gmd:processStep>
    <gmd:source>
      <gmd:LI_Source id="HRPT_AVHRR_L1B">
        <gmd:description>
          <gco:CharacterString>
            HRPT is a live data feed as the spacecraft goes over a receiving stations.
          </gco:CharacterString>
        </gmd:description>
        <gmd:sourceCitation></gmd:sourceCitation>
        <gmd:sourceExtent>
          <gmd:EX_Extent>
            <gmd:temporalElement>
              <gmd:EX_TemporalExtent>
                <gmd:extent>
                  <gml:TimePeriod gml:id="tp_1030059.81238">
                    <gml:beginPosition>2003-11-10</gml:beginPosition>
                    <gml:endPosition/>
                  </gml:TimePeriod>
                </gmd:extent>
              </gmd:EX_TemporalExtent>
            </gmd:temporalElement>
          </gmd:EX_Extent>
        </gmd:sourceExtent>
        <gmd:sourceStep xlink:href="#121"/>
      </gmd:LI_Source>
    </gmd:source>
    <gmd:source>
      <gmd:LI_Source id="HRPT_AVHRR_L0">
        ...
      </gmd:LI_Source>
    </gmd:source>
    <gmd:source>...</gmd:source>
    <gmd:source>...</gmd:source>
    <gmd:source>...</gmd:source>
    <gmd:source>...</gmd:source>
  </gmd:LI_Lineage>
</gmd:lineage>
```

Lineage as a Component

The same inputs or processing chains are many times used to create a series of results or files. In some cases, it may make sense to describe these elements once and access them as components. There are many options for the granularity of this approach and the choice depends on how often the data sets and processing systems involved change. For example, satellite data processing involves at least four kinds of data in processing systems: other products, ancillary data, auxiliary data, and lookup tables. These inputs change on different schedules, e.g. input satellite products may change every cycle while global digital elevation models may never change. Repeating complete source information that rarely changes increases the size of the metadata records and can be distracting. Referencing these quasi-static sources instead of including them simplifies the metadata and focuses it on lineage items that change.

This example shows a schematic processStep that includes one of each source type (complete contents are only shown for the first source):

```

<gmd:lineage>
  <gmd:LI_Lineage>
    <gmd:processStep>
      <gmi:LE_ProcessStep>
        <gmd:description>Brief Text</gmd:description>
        <gmd:source>
          <gmi:LE_Source uuid="uniqueIdentifierForSource">
            <gmd:sourceCitation>
              <gmd:CI_Citation>
                <gmd:title>
                  <gmx:FileName src="InputFileURI">Unique product file name</gmx:FileName>
                </gmd:title>
                <gmd:date>
                  <gmd:CI_Date>
                    <gmd:date/>
                    <gmd:dateType>
                      <gmd:CI_DateTypeCode codeList="" codeListValue="creation">creation</gmd:CI_DateTypeCode>
                    </gmd:dateType>
                  </gmd:CI_Date>
                </gmd:date>
              </gmd:CI_Citation>
            </gmd:sourceCitation>
          </gmi:LE_Source>
        </gmd:source>
        <gmd:source uuid="uniqueIdentifierForSource">Ancillary Data</gmd:source>
        <gmd:source uuid="uniqueIdentifierForSource">Auxiliary Data</gmd:source>
        <gmd:source uuid="uniqueIdentifierForSource">Lookup Table</gmd:source>
        <gmi:processingInformation>
          <gmi:algorithm>
            <gmi:description>Brief Text</gmi:description>
            <gmi:citation>Citation</gmi:citation>
          </gmi:algorithm>
          <gmi:softwareReference>Citation</gmi:softwareReference>
          <gmi:documentation>Citation</gmi:documentation>
        </gmi:processingInformation>
        <gmi:output uuid="uniqueIdentifierForSource">Output Product</gmi:output>
      </gmi:LE_ProcessStep>
    </gmd:processStep>
  </gmd:LI_Lineage>
</gmd:lineage>

```

The example above shows how internal references can be used to simplify the XML representation of the lineage. If the inputs to the process step are uniquely identified, and a service exists that can provide the XML for a particular component, the sources can be referenced as external components to make a more compact lineage description:

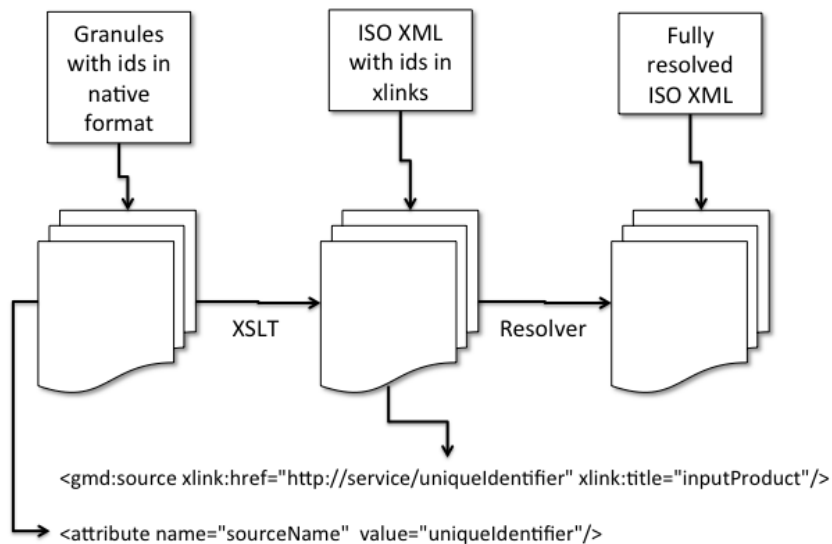
```

<gmd:lineage>
  <gmd:LI_Lineage>
    <gmd:processStep>
      <gmi:LE_ProcessStep>
        <gmd:description>Brief Text</gmd:description>
        <gmd:source xlink:href="http://service/uniqueIdentifier" xlink:title="inputProduct"/>
        <gmd:source xlink:href="http://service/uniqueIdentifier" xlink:title="inputAncillaryData"/>
        <gmd:source xlink:href="http://service/uniqueIdentifier" xlink:title="inputAuxiliaryData"/>
        <gmd:source xlink:href="http://service/uniqueIdentifier" xlink:title="inputLookupTable"/>
        <gmi:processingInformation xlink:href="http://componentService/uniqueIdentifier" xlink:title="Standard
Processing Information"/>
        <gmi:output xlink:href="http://service/uniqueIdentifier" xlink:title="outputProduct"/>
      </gmi:LE_ProcessStep>
    </gmd:processStep>
  </gmd:LI_Lineage>
</gmd:lineage>

```

For example, the lineage description shown above is a component with a unique identifier (UUID) of 95BD4CCC-D27D-8DE4-E040-0AC8C5BB43B64. it can be accessed using <http://www.ngdc.noaa.gov/docucomp/95BD4CCC-D27D-8DE4-E040-0AC8C5BB43B64>.

Lineage in Granules



Granule Lineage

Series or collection metadata provides an overview of the lineage for a series by identifying sources and processing systems used to produce the entire series. Many of these inputs and processing systems change during the life time of a series. Dataset or granule metadata provides information about the specific source granules used to produce a specific granule in the series.

There are several important characteristics of granule lineage that must be considered in decisions about how it should be represented. First, it must be written in the native format of the granule in order to be accessible to clients that are reading the data. Second, for many users, the actual lineage details are less important than easily accessible information about when lineage elements change. In other words, users do not need to be told explicitly (i.e. over and over) what auxiliary dataset is used as a global temperature climatology in each granule, but they definitely need to know when that input changes.

The component approach described above can be useful for addressing these specific requirements. For example, lineage is represented in the netCDF Climate-Forecast conventions with a single attribute called history. The DART lineage shown above could be referenced using this single attribute:

```
<nc:attribute name="history" value="95BD4CCC-D27D-8DE4-E040-0AC8C5BB43B64">
```

When the netCDF documentation (ncml) is translated to ISO this processing would be represented as: `<gmd:lineage xlink:href="http://www.ngdc.noaa.gov/docucomp/95BD4CCC-D27D-8DE4-E040-0AC8C5BB43B64" xlink:title="Dart Bouy D165 Processing"/>` which would unambiguously identify the processing without including excessive redundant information in each granule. To see if lineage was constant, the user would only need to check whether this id changed. When a user needed the complete lineage description, the `xlink:href` would be resolved.

This single identifier approach could work in simple cases, but most real-world situations need more granularity. This can be achieved by providing identifiers for each source in a lineage group:

```
<nc:group name="lineageInformation">
  <attribute name="processStep" value="uniqueIdentifier"/>
  <attribute name="inputProduct" value="uniqueIdentifier"/>
  <attribute name="inputAncillaryData" value="uniqueIdentifier"/>
  <attribute name="inputAuxiliaryData" value="uniqueIdentifier"/>
  <attribute name="inputAuxiliaryData" value="inputLookupTable"/>
</nc:variable>
```

The identifiers in this case could be UUIDs or unique filenames. They would be translated into xLink references (as shown above) in the compact ISO representation of this granule metadata and resolved into complete sources in the complete ISO metadata (see Figure). A user would need to check all source and processing identifiers to determine if or how the lineage for a specific granule had changed.