International Metadata Standards and Enterprise Data Quality Metadata Systems

IN027 - December 15, 2016

Ted Habermann
Director of Earth Science
The HDF Group
thabermann@hdfgroup.org

This work was supported by NASA/GSFC under Raytheon Co. contract number NNG15HZ39C
The Big Picture

ISO 19157 is a conceptual model of data quality metadata that was recently approved as an international standard. It combines three older standards into a unified model for describing data quality.

Many of the principle elements of this conceptual model are abstract, and can be implemented in several ways.
The Big Picture

ISO 19157 is a conceptual model of data quality metadata that was recently approved as an international standard. It combines three older standards into a unified model for describing data quality.

Many of the principle elements of this conceptual model are abstract, they can be implemented in several ways.

When only the abstract concepts are considered, the model is very simple.
Enterprise Systems?
Stand Alone Quality Reports

“There are papers and web pages that describe the quality of my data.”

Papers and reports that describe data quality are StandAloneReports. Metadata can include brief descriptions of the results (abstracts) and references to any number of these (citations).

Abstract: The fire training-set may also have been biased against savanna and savanna woodland fires since their detection is more difficult than in humid, forest environments with cool background temperatures [Malingreau, 1990]. There may, therefore, be an under-sampling of fires in these warmer background environments.

Data Usage (19115-1)

“Users increase our understanding of data quality. We need to keep them in the loop.”

MD_Usage

+ specificUsage : CharacterString
+ usageDateTime [0..1] : DateTime
+ userDeterminedLimitations [0..1] : CharacterString
+ userContactInfo [1..*] : CI_ResponsibleParty
+ response [0..*] : CharacterString
+ additionalDocumentation [0..*] : CI_Citation
+ identifiedIssues [0..1] : CI_Citation

DOI

View the 06_L2 HISTORY page for details on correction.
What is a Data Quality Element?

![Diagram showing a Data Quality Element with QA_PercentMissingData, Measure, Method, and Result.]

- **QA_PercentMissingData**
- **Measure**
- **Method**
- **Result**

**Number of Pixels with Missing Flags**

\[
\frac{\text{Number of Pixels with Missing Flags}}{\text{Total Number of Pixels}} = 15\%
\]
What Are Quality Measures?

“My metadata already include data quality measures.”

NASA EOSDIS metadata includes two types of quality measures.

- **QA_Stats**
  - QAPercentMissingData – Granule level % missing data for individual parameters within a granule.
  - QAPercentOutOfBoundsData – Granule level % out of bounds repeated for individual parameters within a granule.

- **QA_Flags**
  - QAFlags – The name of the geophysical parameter expressed in the data as well as associated quality flags and quality status.
    - AutomaticQualityFlag – The granule level flag applying generally to the granule and specifically to parameters at the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developer and documented in the Quality Flag Explanation.
    - OperationalQualityFlag – The granule level flag applying both generally to a granule and specifically to parameters at the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developer and documented in the Operational Quality Flag Explanation.
    - ScienceQualityFlag – Granule level flag applying to a granule, and specifically to parameters. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developer and documented in the Science Quality Flag Explanation.
What Are Quality Measures?

“I use consistent Quality Measures across many products.”

**QAStats** – **Standard measures for all products**

**QAPercentMissingData** - Granule level % missing data. This attribute can be repeated for individual parameters within a granule.

**QAPercentOutOfBoundsData** – Granule level % out of bounds data. This attribute can be repeated for individual parameters within a granule.

**QAPercentInterpolatedData** – Granule level % interpolated data. This attribute can be repeated for individual parameters within a granule.

**QAPercentCloudCover** – This attribute is used to characterize the cloud cover amount of a granule. This attribute may be repeated for individual parameters within a granule. (Note - there may be more than one way to define a cloud or it's effects within a product containing several parameters; i.e. this attribute may be parameter specific)
What Are Quality Measures?

“I use consistent types of Quality Measure across many products.”

**QAFlags** – Classes of quality measures with product specific implementations

**AutomaticQualityFlag** – The granule level flag applying generally to the granule and specifically to parameters the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developer and documented in the Quality Flag Explanation.

**AutomaticQualityFlagExplanation** – A text explanation of the criteria used to set automatic quality flag, including thresholds or other criteria.

**OperationalQualityFlag** – The granule level flag applying both generally to a granule and specifically to parameters at the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developers and documented in the Operational Quality Flag Explanation.

**OperationalQualityFlagExplanation** – A text explanation of the criteria used to set operational quality flag; including thresholds or other criteria.

**ScienceQualityFlag** – Granule level flag applying to a granule, and specifically to parameters. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developers and documented in the Science Quality Flag Explanation.

**ScienceQualityFlagExplanation** – A text explanation of the criteria used to set science quality flag; including thresholds or other criteria.
Data Quality Measures

“My data quality measures are consistently described in a database.”

ISO 19157 includes a DQ_MeasureReference designed to provide a connection to a detailed description of the quality measure.
Data Quality Measures

“I need to clearly and consistently explain how I measure quality.”

The ISO model for quality measures includes identifiers, definitions, descriptions, references and illustrations.

<table>
<thead>
<tr>
<th>Line</th>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name</td>
<td>number of invalid self-overlap errors</td>
</tr>
<tr>
<td>2</td>
<td>Alias</td>
<td>kickbacks</td>
</tr>
<tr>
<td>3</td>
<td>Element name</td>
<td>topological consistency</td>
</tr>
<tr>
<td>4</td>
<td>Basic measure</td>
<td>error count</td>
</tr>
<tr>
<td>5</td>
<td>Definition</td>
<td>count of all items in the data that illegally self overlap</td>
</tr>
<tr>
<td>6</td>
<td>Description</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>Parameter</td>
<td>–</td>
</tr>
<tr>
<td>8</td>
<td>Value type</td>
<td>Integer</td>
</tr>
<tr>
<td>9</td>
<td>Value structure</td>
<td>–</td>
</tr>
<tr>
<td>10</td>
<td>Source reference</td>
<td>–</td>
</tr>
<tr>
<td>11</td>
<td>Example</td>
<td><img src="image" alt="Example diagram" /></td>
</tr>
<tr>
<td>12</td>
<td>Identifier</td>
<td>27</td>
</tr>
</tbody>
</table>

Table D.28 — Number of invalid self-overlap errors

<table>
<thead>
<tr>
<th>Line</th>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name</td>
<td>mean value of positional uncertainties excluding outliers (2D)</td>
</tr>
<tr>
<td>2</td>
<td>Alias</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>Element name</td>
<td>absolute or external accuracy</td>
</tr>
<tr>
<td>4</td>
<td>Basic measure</td>
<td>not applicable</td>
</tr>
<tr>
<td>5</td>
<td>Definition</td>
<td>for a set of points where the distance does not exceed a defined threshold, the arithmetical average of distances between their measured positions and what is considered as the corresponding true positions</td>
</tr>
<tr>
<td>6</td>
<td>Description</td>
<td>For a number of points ( N ), the measured positions are given as ( x_{1m}, y_{1m}, z_{1m} ) and ( x_{2m}, y_{2m}, z_{2m} ). The coordinates depend on the dimension in which the position of the point is measured. A corresponding set of coordinates ( x_{2m}, y_{2m}, z_{2m} ) and ( x_{1m}, y_{1m}, z_{1m} ) are considered to represent the true positions. All positional uncertainties above a defined threshold ( \delta_{\text{th}} ) are then removed from the set. The positional uncertainties are calculated as</td>
</tr>
<tr>
<td>7</td>
<td>Parameter</td>
<td>Name: ( \delta_{\text{th}} )</td>
</tr>
<tr>
<td>8</td>
<td>Value type</td>
<td>Measure</td>
</tr>
<tr>
<td>9</td>
<td>Value structure</td>
<td>–</td>
</tr>
<tr>
<td>10</td>
<td>Source reference</td>
<td>–</td>
</tr>
<tr>
<td>11</td>
<td>Example</td>
<td>–</td>
</tr>
<tr>
<td>12</td>
<td>Identifier</td>
<td>29</td>
</tr>
</tbody>
</table>
Enterprise Systems?

Measures  Methods  Reports / Documents
“There are papers and web pages that describe the quality of my data.”

“Users increase our understanding of data quality. We need to keep them in the loop.”

“I use consistent types of Quality Measure across many products.”

“My metadata currently includes descriptions of the quality of my data.”

“My data quality information exists in databases or web services.”

“The quality of my data vary in time and space and different parameters have different quality measures and results.”

“I use consistent Quality Measures across many products.”

“I need to clearly and consistently explain how I measure quality.”
Acknowledgements

This work was partially supported by contract number NNG15HZ39C from NASA.

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of NASA or The HDF Group.