Describing the quality of research datasets across disciplines: a comparative study

Tiffany C. Chao  tchao @ illinois.edu

Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science, University of Illinois

OVERVIEW
Descriptions of data quality play an integral role in data reuse determinations by potential users. However, quality characterizations vary across different research cultures where formalized criteria may be limited or may not be sufficient for curation purposes.1

Addressed in this poster is a preliminary examination of research dataset records from three sub-disciplines of Earth science that centers on:

- how quality is described by scientists for research datasets
- what patterns in quality description emerge across different sub-disciplinary fields

ASSESSMENT OF QUALITY DESCRIPTIONS

SAMPLE SCOPE: (3) sub-disciplines of Earth science from the Global Change Master Directory (GCMD) (http://gcmd.nasa.gov):

- Geochemistry
- Population science
- Atmospheric science

SOURCE: Dataset metadata records from GCMD (collected in Fall 2012); records follow the DIF format (Directory Interchange format) which includes a specific field for quality.

APPROACH: Extracted available descriptions from the <quality> field for each dataset record and assigned a category based on prescribed DIF “quality” definitions. Reviewed initial categories for emergent topics and recoded descriptions with new list. Only one category was assigned to each description.

The <quality> field allows the author to provide information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata.2

This information may include:
- Indicators of data quality or quality flags
- Established quality control mechanisms
- Established quantitative quality measurements
- Recognized or potential problems with data quality

The <quality> field is highly recommended but not required for the DIF record

<table>
<thead>
<tr>
<th>Observed &lt;quality&gt; categories in DIF records across all three sub-discipline areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements: for funding or external support</td>
</tr>
<tr>
<td>Data description: details about the provided data (i.e. format, variable names, study context, etc.)</td>
</tr>
<tr>
<td>Disclaimer: warning about using the data (i.e. data provider not liable for accuracy)</td>
</tr>
<tr>
<td>Method: details about how the data were collected and analyzed</td>
</tr>
<tr>
<td>Problem: statements related to the completeness of data provided (i.e. gaps in data collection, potential data corruption, etc.)</td>
</tr>
<tr>
<td>Quality assurance procedures: identified persons or procedures used for quality checks and review of data</td>
</tr>
<tr>
<td>Quality indicator: usefulness of provided data; marked status of provided data (i.e. “as is”, “approximate values”)</td>
</tr>
<tr>
<td>Reference: to an external source of information (i.e. data provider contact information; physical location of data; URL, etc.)</td>
</tr>
</tbody>
</table>

CONCLUSIONS
The quality of data can be described in a number of ways. In applying the DIF field definitions for quality to available records, different interpretations emerged resulting in additional categories. To a certain extent, some of the observed categories may actually fit in with existing definitions while others may not be related to quality at all but to other aspects of the data.

The comparison of data quality descriptions between geochemistry, population science, and atmospheric science revealed similar types of information publically conveyed, with a focus on how the data were generated (Method) and what additional evidence was available (Reference) to potentially establish quality. The amount of detail presented did vary within each sub-discipline area where no consistent patterns were observed.

Directions for future work consider how informative current data quality descriptions actually are as assessed by research scholars and curation professionals and how these descriptions could be improved. Such assessments have implications for the level of curatorial service allocated to the dataset in order to support future access and reuse.

REFERENCES:

The top three observed quality categories for datasets in each sub-discipline along with examples from the record descriptions are listed below. The number after each category indicates the percentage of records that fall under that particular category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Description</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geochemistry</td>
<td></td>
<td>48%</td>
</tr>
<tr>
<td>Geographic science</td>
<td></td>
<td>25.85%</td>
</tr>
<tr>
<td>Atmospheric science</td>
<td></td>
<td>15.79%</td>
</tr>
</tbody>
</table>

The Problems related to provided data varied in the level of detail describing the nature of the issue. For instance, some statements about the atmospheric science data discussed specific areas where the data provided may be compromised or unreliable whereas other statements were more general.

These described problems were also examined in relation to data types and whether similar issues were identified within each of the sub-disciplinary fields. Though three data types, or formats, were found in common (ASCII, PDF, and shapefiles), descriptions of quality did not discuss problems but spanned the other observed quality categories.

The <quality> field allows the author to provide information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata. This information may include:

- Indicators of data quality or quality flags
- Established quality control mechanisms
- Established quantitative quality measurements
- Recognized or potential problems with data quality

The <quality> field is highly recommended but not required for the DIF record.

The top three observed quality categories for datasets in each sub-discipline along with examples from the record descriptions are listed below. The number after each category indicates the percentage of records that fall under that particular category.