Towards a Cross-Disciplinary Notion of Data Level in Data Curation

Allen H. Renear1, Molly Dolan2, Kevin Trainor3, Melissa H. Cragin4
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street, MC-493, Champaign, IL 61820-6211 USA
1renear@illinois.edu, 2madolan@illinois.edu, 3trainor1@illinois.edu, 4cragin@illinois.edu

Data Levels in Earth Science and Textual Criticism

- Originally focused on scientific data, the data curation community is now engaged with humanities data as well.
- Sharing concepts and terminology across domains will be valuable for both:
  i. the practice of data curation, and
  ii. the education of data curation professionals.
- Can these distinct domains support shared frameworks of common concepts?
- As an exercise in conceptual alignment, we compare the widely used NASA data level categories for remote sensing data with traditional notions of scholarly transcription and editing found in textual criticism or textual philology.
- “Data level” categorizes data with respect to the extent to which it is “raw” or “processed”.

Data Levels in Earth Science and Textual Criticism

<table>
<thead>
<tr>
<th>NASA Data Levels</th>
<th>TextCrit Data Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Level 0</td>
</tr>
<tr>
<td>“…unprocessed instrument data at full resolutions.”</td>
<td>Unprocessed text images.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 1A</th>
<th>Level 1A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed text images, annotated with the identification of the hardware and software used, any configuration or calibration information, the time and place of the scanning, and organization or persons conducting the imaging, and a (non-descriptive) identification of the object imaged.</td>
<td>Administrative and technical metadata elements for raster images (NARA, 2004)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Derived environmental variables (e.g., ocean wave height, soil moisture, ice concentration) at the same resolution and location as the Level 1 source data.”</td>
<td>“Variables mapped on uniform spacetime grid scales, usually with some completeness and consistency properties (e.g., missing points interpolated, complete regions mosaicked together from multiple orbits).”</td>
</tr>
</tbody>
</table>

Level 3

- Representation of textual content and structure mapped on to (perhaps multiple) carriers with described structure (e.g., physical bibliography), interpolation of missing text (lacunae).

Level 4

- Textual history including scribal transmission, seriation, intended but unrealized texts, etc. Possibly also person, name, and date disambiguation.

Questions for Further Discussion

- Are there fundamental differences between cultural and scientific data that will bear on the characterization of data level?
- What role does human judgment and intuition play in moving from one data level to another?
- Is this role the same in the sciences and the humanities?
- How does the intentionality, the aboutness, of cultural artifacts fit into the concept of data levels?
- What is the effect of one discipline’s theory being another discipline’s data?
- Is a scholarly edition data for a literary critic, but theory for a textual philologist?
- These are operational definitions; how can we characterize data levels conceptually?

Acknowledgements

We have benefited from discussions with Carole L. Palmer, John MacMullen, and Trevor Muñoz as well as from discussions at the e-Research Roundtable, Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. This work is supported by IMLS grant RE-05-08-0062-08.

References