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

Allen H. Renear[1], Molly Dolan[2], Kevin Trainor [3], Melissa H. Cragin[4]
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
[1]renear@illinois.edu, [2]madolan@illinois.edu, [3]trainor1@illinois.edu, [4]cragin@illinois.edu

CIRSS
INSTITUTE of Museum and Library SERVICES

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".

NASA Data Levels	TextCrit Data Levels
Level 0	
"…unprocessed instrument data at full resolutions."	Unprocessed text images.
	Bitmap images of textual documents
Level 1A	
"…unprocessed instrument data at full resolution, time referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters (i.e., platform ephemeris) computed and appended but not applied to the Level 0 data."	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.
	Administrative and technical metadata elements for raster images (NARA, 2004)
Level 2	
"Derived environmental variables (e.g., ocean wave height. soil moisture, ice concentration) at the same resolution and location as the Level 1 source data."	Derived representation of text content and structure, mapped to locations in the Level 1 source data.
	Transcriptions of Level 1 text images with accompanying markup indicating textual components (XML/TEI documents)
Level 3	
"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)."	Representation of textual content and structure mapped on to (perhaps multiple) carriers with described structure (e.g, physical bibliography), interpolation of missing text (*lacunae*).
	Locating the digital text in a larger bibliographic space of physical objects and editions
Level 4	
"Model output or results from analyses of lower-level data (i.e., variables that were not measured by the instruments but instead are derived from these measurements)."	Textual history including scribal transmission, seriation, intended but unrealized texts, etc. Possibly also person, name, and date disambiguation.
	•Correction (*emendation*) of textual errors •Interpolation of missing manuscripts •Determination of order of composition (*seriation*) •Tree of manuscript transmission

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? (e.g., a scholarly edition is 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

Bose, R. and Frew, J. (2005). Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.* 37(1), 1-28.
Cragin, M.H., Heidorn, P.B., Palmer, C.L., Smith, L.C. (2007). An educational program on data curation. *Poster, Science and Technology Section of the annual American Library Association conference. Washington, D.C., June 25, 2007.*
Crane, G., Babeu, A., & Bamman, D. (2007). eScience and the humanities. *International Journal of Digital Libraries*, 7.
Freire, J., Koop, D., Santos, E., Silva, C.T. (2008). Provenance for computational tasks: a survey. *Computing in Science & Engineering*, 10(3), 11-21.
NARA (2004). *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files.* Written by Steven Puglia, Jeffrey Reed, and Erin Rhodes. U.S. National Archives and Records Administration.
NASA. (1986). Earth observing system. Data and information system. Volume 2A: Report of the EOS Data Panel. NASA Technical Memorandum 87777.
Rusbridge, C. (2007). Create, curate, re-use: the expanding life course of digital research data. Paper, *EDUCAUSE Australasia 2007.*
Swan, A. & Brown, S. (2008). The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs.
TEI (2008). *TEI P5: Guidelines for Text Encoding and Interchange*. TEI Consortium, Oxford, Providence, Charlottesville, Nancy.