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

Allen H. Renear\(^1\), Molly Dolan\(^2\), Kevin Trainor \(^3\)

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

Acknowledgements: We have benefited from discussions with Melissa H. Cragin and Carole L. Palmer.
Data Curation

• Data curation is concerned with the fundamental principles and best practices for managing the entire lifecycle of data: creation, management, exploitation, enhancement, and preservation (cf Cragin, 2007, Rusbridge, 2007; others).
DCEP

• The GSLIS Data Curation Education Program (DCEP)

• A specialization within GSLIS masters curriculum

• 2008 IMLS grant: to extend the DCEP program to include the humanities as well as scientific data.
DCEP-H Ideology

• Everyone agrees:
  the digital context requires better-theorized curatorial practices.
  – wrt to authenticity, provenance, authoritative reference, and annotation, etc.

• However, the humanities disciplines have already evolved sophisticated curation practices and theories

• And in the last 50 years these have been refined and extended, already, to support computation and digital storage.

• This rich tradition must inform any development of a humanities data curation curriculum.
Cross-disciplinary curatorial concepts

• To what extent can shared frameworks of common curatorial concepts support best practices, current and future, across the disciplines?
The Value of Shared Concepts

• Data curation services may be provided by a single organizational unit serving multiple disciplines, with a single set of repository tools, and a single set of policies (Swan & Brown, 2008).
  – Unnecessary diversity will create a burden for design, implementation, documentation, and training.

• Concealing similarities makes it harder to recognize opportunities for re-use of successful strategies in different domains.

• An integrated curriculum has obvious pedagogical advantages.

• Most importantly:

  the more the data curation curriculum reflects commonalities across superficially disparate data practices, the more robust and effective the theory and techniques we teach in our curriculum are likely to be.
First target concept: Data Levels

the categorization of a data product with respect to the extent to which it is “raw” or “processed”, and the documentation of the transformation of data from one stage to another,

[related to what the data curation community now calls *provenance*, or more specifically: *lineage* (Bose and Frew, 2005)].
First conceptual alignment exercise

- a comparison of:
  - the widely used *NASA data level categories* for remote sensing data (NASA, 1986)
  - the equally important traditional concepts of editorial intervention in *textual criticism*.
(What we are really talking about here)

- data / evidence / observation …

vs

- theory / inference / interpretation …
Textual Criticism

- the study of textual materials to determine fundamental facts about their nature and relationships (*and the textual intentions of their authors*)
  - what the “literal” text is
  - what was copied from what, what was written first, etc.
  - what mistakes were made, and why etc.

- aka *textual bibliography, textual philology, higher/lower criticism, …* (but not literary criticism)

- Highly evolved, broad influence.

- all traditional cultural knowledge sits on a foundation of textual criticism (and science too according to textual bibliographer Thomas Tanselle)

- Text critical concerns have been enormously influential in the last(first) 60 years of humanities computing
Level 0

**NASA Level 0:** “…unprocessed instrument data at full resolutions.”

**TextCrit Level 0:** unprocessed text images.

Images of textual documents (codex or manuscript pages for instance) recorded in raster formats (bitmaps) and prior to any additional processing, such as format conversion or compression.

- At the original scanning resolution, so may contain details that can be lost when smaller more manageable versions are generated.
- Though features apparent only after processing may not be evident when rendered with standard viewers.
- Usually inappropriate for digital research as they are large, often in inconvenient formats, and do not make textual content computationally available.
- Nevertheless they are arguably the *evidentiary foundation* for all further scholarship.
Level 1

• NASA 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.

• TextCrit Level 1A: Unprocessed text images, annotated with the identification of hardware and software used, configuration or calibration, scanning time and place, and organization or persons conducting the imaging, and a [non-descriptive] identification of the object imaged.
  – Here information about the scanning process is included, as well as a specification of what was scanned.
    • “objective” administrative and technical metadata elements for raster images (NARA, 2004).
  – And non-descriptive identification of the object being scanned.
    • the combination of an agency and a registrar’s accession number would be an example of a relatively non-descriptive identification; characterization of the document as written by a particular author would not — involving too much interpretation and inference.
Level 2

- **NASA 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.

- **TextCrit Level 2**: Derived representation of text content and structure, mapped to locations in the Level 1 source data.

  - e.g. transcriptions of Level 1 text images with markup indicating non-interpretative textual components.
    - perhaps a TEI document intended for broad use by diverse scholars?
    - but without (e.g.)
      - additional annotation as to relationships to other texts,
      - other physical instantiations of this text, or other raster files carrying this text,
      - or the identification of mentioned persons, places, times, etc.
  - Not only a close to match to Level 2 in nature, but plays the same role in much scholarship that NASA Level 2 data plays in earth science:
    - because processed enough to be useful, but without information that might be considered interpretation by scholars using them.
Aside on Interpretation

- Many difficult decisions arise at this point:
  - What counts as a data feature to be recorded?
    - Are purely typographic variations distinguished?
    - Are (apparently) non-meaningful design ornaments to be recorded?
    - The literatures of textual criticism, bibliography, documentary editing and diplomastics provide extensive discussions of possible transcription policies.
- Another issue is, obviously, what counts as interpretation.
  - Printed glyphs and characters might be considered identifiable with (usually) relatively little interpretation.
  - Text components that can be identified with relatively little interpretation might be chapter and paragraph boundaries, extracts, stanzas,
    - and other things for which there is reasonably non-controversial visual evidence
  - while the distinction of place names and person names, or the identification of referenced persons, places, and things are probably annotations too inferential for Level 2
Level 3

• **NASA 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).

• **TextCrit Level 3:** Representation of textual content and structure mapped on to (perhaps multiple) carriers with described structure (e.g, physical bibliography), expansion of abbreviations, interpolation of missing text.

• Locating the digital text in a larger bibliographic space of physical objects and editions corresponds well to mapping data to a “uniform spacetime grid”.
  – Allows several raster images and corresponding TEI documents to be mapped to the same book, or the same edition, perhaps “mosaicking” several XML/TEI documents together.
  – The standard text critical practice of supplying missing text where there are “lacunae” caused by illegibility, physical damage, etc. corresponds well to the interpolation of missing points.
    • [but are these two interventions are really at the same epistemic level of processing and inference?]
Level 4

- **NASA 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).

- **TextCrit Level 4:** Textual history including scribal transmission, seriation, intended but unrealized texts, etc. Possibly also person, name, and data disambiguation.

- This level would contain the results of other aspects of traditional textual criticism,
  - such as the correction (*emendation*) of textual errors made by the original authors, scribes, or compositors,
  - the interpolation of missing manuscripts, the determination of order of composition (seriation),
  - or the establishment of an intended text that was never realized due to compositor errors.

- None of these things are “directly measured”, but they are derived from what is measured.
End
• One is differing notions across different humanities disciplines of what is data and what is theory.

• An “established” text is data for a literary critic or cultural historian, but an end product and theoretical achievement for a textual philologist.

• Another possible difference is the frequent role of human judgment and intuition in moving from one data level to another — even when algorithmic techniques are employed (in computational seriation or stemmatics for instance) human judgment remains salient. And finally, there is the intentionality, the *aboutness*, of humanities data: an inscription not only has a causal history, but it is intended to have meaning as well.

• Whether these are deep differences between science and humanities data, or only general differences in emphasis (or entirely illusory) remains to be seen. As we have noted, the exercise has been set — let the discussion begin.
References


