TOWARDS STRATEGIC READING? GRAPHICAL MAPS AND RENDITIONS OF TEI DATA

At JATS-Con 2013/4 [Renear 2014], Allen Renear made the fascinating suggestion that one benefit of descriptive encoding will be in "strategic reading", which he described as a process of reading, assessment and engagement that happens at a "higher level" than simply reading. It involves a kind of meta-level assessment and integration that happens along with the task of "reading" per se (discursive interpretation of written language): relating data, metadata, and information that is acquired heuristically and interpretively, focused not on content so much as on structures such as equations, propositions, formal objects.

We can take this notion quite literally and wonder whether an actual rendition or depiction of a document "from a distance" and focusing on structures as much as on content, would be a worthwhile kind of experiment. If we can't automate strategic reading, maybe we can facilitate it, and certainly visualization is all the rage right now. Also, we know at least in principle how to do this -- already ten years ago at least two TEI practitioners were demonstrating TEI to SVG via XSLT (see [Cayless 2005], [Piez 2005]).

Then too, we have better platforms now than we had then, and we have needed to try some things (and still do).

"Drawing" the data with SVG is certainly one of the more tantalizing applications of XSLT. Among the data sets I am currently looking at is the journal content of DHQ (Digital Humanities Quarterly at http:digitalhumanities.org/dhq/), which is encoded in a variant of TEI (lightly extended). Of course an interface to browse DHQ also hints at the application of (the same or similar) XSLT to any TEI to produce similar diagrams, or (better) diagrams in forms reflecting local (not only TEI) markup semantics.

See the attachment, which includes (1) a query results page, with (2) several linked pages for individual articles. The query results page presents one set of visualizations of the documents' XML structure, a "bubble diagram" (in cross-section). From there, one may also link to a "map" of each (any) document to show its hierarchical structure in diagrammatic form, alongside a reduced (greatly minimized) text (HTML) rendition of the article. (Currently this is fairly roughly sketched; much more can be done to enhance and elaborate.)

In the demonstration I have built, the transformation pipelines are executed in XSLT from an XQuery database (BaseX), which is able to provide a web-based interface to the document repository. This gives me the flexibility needed to (for example) provide querying the documents in order to isolate and select interesting examples for rendering. (The fact that the query results are shown in graphic form is incidental to this functionality, but they go well together.) So, for example, you can query for "all the figures in documents that have five figures or more". In the

diagrammatic display of the documents returned, figures (or whatever the query return set may happen to be) are highlighted.

Once able to graph out either an element-based (tree-shaped) network, or a set of measured spans (typically measuring document contents), we can extend the logic into the illustration of the features and structures in known document formats such as TEI. First we draw a representation of the tree (elements and/or range spans), then we start ornamenting and labeling it. Labels can include the extraction of data and metadata from the document dynamically.

One of the interesting problems we run across here is that display of the tree (the element "containment graph") in its "correct proportions" is actually harder than it seems it ought to be. On the one hand, this implies an annotated structural diagram showing simply the relations of elements; but on the other, it suggests that relative proportions of different "sizes" or "extents" of content also be indicated somehow. It is surprisingly and interestingly difficult to do both these things at once, since each implies a different rendering logic. To analyze "lengths", document content must be measured; to depict hierarchical relations of "containment", the element tree must be measured. The problem becomes more tractable if it is split into layers: one process (XSLT transformation) can capture an abstract document map (in a hierarchy with extents noted), and then in another XSLT, we process the map for display in SVG (determining at that point the particulars of the layout).

Technical issues aside, however, the main question remains as it was ten years ago: what we learn from this, why are we doing it. This question is not answered by the arguably "denaturalizing effect" this kind of treatment seems to have on the documents. As readers, our attention is directed not at reading but at discovering interesting features revealed by patterns in markup (it is curious to see which documents have exactly five 'div' elements in the body); thus documents are more interesting as examples of genre or "subtype", no longer in their contents, what they purport to "be about" and presumably what their authors were most concerned with in composition. Consequently it might be argued that such scans would tend to distract from strategic reading, not support it (however they might be interesting to us for themselves). And while it seems that graphic representations of structures in an XML document should be self-evidently interesting (having put the elements there surely we want to see them) -- on the other hand (and this is entirely subjective) it can also be somewhat unsettling.

One set of questions regards how the "phenomenology" of the document or indeed the journal content changes when we approach it so differently. Documents, so reduced, lose whatever distinctive "voices" they have; and the distinctions they gain seem, if not arbitrary, then somehow clinical. For example, it's quite easy to see how big a bibliography is or whether and how many figures there are. It's less easy to see if an argument is well made. We can have citations pop out, while the motive of the citation recedes, how the article engages with the source it cites. Instead, these diagrams dramatize the rather mechanical and reductive nature of the XML

encoding. We see the "clockwork" of the document as a produced artifact, but nothing about its own topic or concern. It is not altogether clear whether this "deformation" of the text [McGann 2004] is a revelation, or only a distortion.

Then too on a level up (speaking as a sometime writer for DHQ) I'm not sure this doesn't in some sense break the implicit contract between me and the journal. Explicitly of course the journal is free to do what it wants (I have given it the rights to that). But it isn't clear what purpose this serves me as a writer, beyond its being an abstract contribution to something larger than me. In other words, what do I (as a reader and writer) need from DHQ, and do experiments like this give them to me?

Who are these diagrams for, and who would find them not only interesting but useful? One solution to this conundrum might be to imagine what authors and readers (not only we toymakers) might want to see. Another might be to wonder whether diagrams of transcriptions and editions, not just of born digital materials, would be of interest. In that context, these experiments are still only groping. But they may be suggestive.

Bibliography

[Cayless 2005] Cayless, Hugh. "DocScapes: Visualizing Document Structures with SVG." ACH/ALLC 2005 (Victoria, BC). See http://tomcat-stable.hcmc.uvic.ca:8080/ach/site/xhtml.xq?id=165.

[McGann 2004] McGann, Jerome. Radiant Textuality: Literature after the World Wide Web. Palgrave, 2004.

[Piez 2005] Piez, Wendell "SVG Visualization of TEI Texts". Poster presented at ACH/ALLC 2005. See http://tomcat-stable.hcmc.uvic.ca:8080/ach/site/xhtml.xq?id=214

[Renear 2014] Renear, Allen. "Strategic Reading, the Future of Scientific Publishing — something for everyone". JATS-Con 2013/14, Bethesda Maryland, April 2014.