Broader, Not Deeper:

The Institutional Imperative for the TEI

Opening

I'm coming to you today as a friend of the TEI, though I hope that what I say will be taken seriously as a recommendation for significant change in the scope and mission of the TEI. I was at first surprised that Peter got in touch with me to speak at the conference. In all honesty, it has been several years since I've had anything to do with the TEI or any type of text encoding. I think most of you know that I've done a lot of text encoding myself, and for some years managed operations that were directly responsible for creating encoded texts or putting them online. But for several years, my work has been very different. Of course the digital library operation at Michigan is still producing and putting online encoded texts, and more now than ever before, but (as we like to say) I'm just a manager and don't do real work anymore. Frankly, I'm about as clueless as could be when it comes to the daily issues that Michigan Digital Library Production Service staff like Paul Schaffner or Chris Powell face. But like so many managers, being removed from the issues that you face not only *doesn't* discourage me from commenting on them, it emboldens me!

Message

I have a relatively simple message for the attendees of the meeting. It is this:

1. By creating a **comprehensive encoding framework**, the TEI effort has been devoted primarily to **serving the needs of individual scholars**, or relatively small collective scholarly efforts. In this, it has been immensely successful and should be recognized as a landmark effort.
2. In order to remain relevant and economically viable, the TEI must shift its focus not entirely, but significantly to the problems of institutional efforts, and primarily

to those of large digital library efforts. Rather than an encoding framework, it must become a **digital library framework**.

[Note that will be talking about TEI-as-guidelines and TEI-as-initiative, and will try to be clear about this.]

TEI as Framework

I believe that there are essentially two types of guidelines for encoding. One is narrow and the expectation is that users will extend or adapt it for their specific purpose. ISO12083 is an excellent example of this type. I was never been able to use ISO12083 without modifications. Even newly minted books or articles strain the boundaries established by that DTD. Is the order of elements in a bibliography citation *always* title-author? Can't a section have more than one heading? ISO12083 typically constrains these sorts of decisions, creating a sort of procrustean encoding bed. The other type of encoding guideline is comprehensive and typically descriptive, and the TEI is one such example. The TEI guidelines take into account different types of documents; they reflect the variety of practice seen in documents that span centuries and wildly various disciplines; they provide both room for and guidance for extension; they balance flexibility with constraint.

The TEI guidelines *excel* as a comprehensive encoding framework. How do they excel? The TEI is thorough, thoughtful, flexible, and provides an open framework. TEI P3 was a monumental accomplishment with only rare and debatable flaw.

Who do the TEI guidelines serve?

1. individual scholar encoding a text
2. a text "cooperative" of several scholars or scholarly craftspersons
3. resource-specific operations such MED or linguistic operation (e.g., MiCASE)
4. hybrid operations like Chadwyck-Healey and Alexander Street Press, which pretend to be scholarly

The TEI has defined its domain as the work of individual scholars, and has underestimated the ways and extent that institutions would use the TEI. Early in its history, the TEI designed the TEI header with a looseness indicative of the designers' expectations that librarians would seldom be creating headers; rather, scholars would. Large-scale text conversion at institutions such as Indiana, Virginia, and Michigan uses the TEI so as to ensure compatibility with the work of scholars and to take advantage of the broad framework created by the TEI. At Michigan alone, this approach has resulted in nearly 20,000 full length monographs being expressed using the TEI, and more than 50,000 smaller article length documents. I suspect that the work of just one or two of these institutions surpasses the combined output of all individual scholars using the TEI guidelines to date.

Institutional "digital library" challenges

If institutions are where it's at, then what are the challenges that an institution faces?

1. Overwhelmingly, in operations like that of the Preservation department at the University of Michigan.
   a. Cornell preservation: math, US social history, New York
   b. Michigan preservation: math, Philippines, US cultural history, "the shelves"; scale is significant (5,000 volumes per year)
   c. LC
   d. Of course FRUS, Wrights, others, but primarily *preservation*
2. EEBO-TCP and similar operations
   a. EEBO-TCP at Michigan and Oxford more than 100 volumes/month fully encoded and transcribed publications from beginning or printing in England and colonies until 18th c.
3. Other challenges, including:
   a. compound documents
   b. digital repositories
   c. archiving
   d. interoperability

e. instructional technology
f. digital library architectures (how the pieces fit together, and how different libraries fit together)

In all of these areas, TEI offers something and has a relationship to the problems being explored. Initiatives like the preservation conversion project or EEBO TCP are essentially text encoding projects, but in other cases the work itself—archives or interoperability—frequently contains encoded texts and communicates between interoperating systems using some variety of XML.

But, as it stands, because of the nature of the TEI's focus, issues related to all of these areas of work will not make their way into current TEI products. Why? I believe it is because of (1) **the focus on the scholar (rather than the institution)** and (2) the preference for description *without simultaneously supporting prescription*.

The Problem of Page Images

I want to share an anecdote that I think highlights this issue of the TEI's relationship to the work of individuals and the work of institutions. I was amused and honored the other day when a colleague forwarded to me an e-mail exchange wherein Lou referred to something he called "JPW's hack," noting "I did a brief informal survey in NY in June 2002, and found that most people I spoke to use what I think of as "JPW's hack" -- add an extra attribute to the <pb> element which specifies the name of the image file corresponding to that page. In my view, there is an outstanding action on the TEI community either to endorse this or to produce a better solution for inclusion in TEI P5 -- suggestions very welcome." I shouldn't get bogged down in the details—Lou does not seem to be able to recall that this hack does *not* involve extra attributes—but here's an interesting case in point where the TEI seems unable to step up to something in a meaningful and timely way. This issue, as Lou accurately describes it and as he has begun to outline in a set of draft recommendations, is almost as old as the TEI itself. (From their first version, the TEI used page-level encoding as an example of the problem of concurrence.) I brought precisely this issue—links to page images—to the TEI's

Technical Review Committee in 1997 (more than 5 years ago), and at that time both Lou and I had the distinct impression that work was under way. As Lou wrote in 1997, “That's my impression too -- that the TRC agreed to a new work item on this topic.”

Why does this anecdote shed light on the issue? Producing a solution for this problem is not currently at the core of the TEI’s mission. Few scholars face the need to provide links to page images, and they rarely encounter this linking/association problem as a need to synchronize parallel instances of a document in two separate systems. [Describe UM’s processes and volume at this point.]

There are two major problems when it comes to the TEI initiative serving the needs of institutions, needs like this one:

1. The Guidelines themselves are very good, but by considering them alone, the TEI is only descriptive and hardly ever prescriptive. Perhaps this is not fair, but in the TEI we lack a set of *application guidelines*, some guidance for the application of TEI by individuals and especially institutions that are frankly not interested in the open-endedness of the TEI.
2. The current **framework** provided by TEI does not extend beyond encoding, and must either encompass the issues facing digital libraries, or find some formal way of meeting up with them.

Those same virtues that I outlined earlier—comprehensiveness, flexibility, description—are sometimes a problem, and especially for a person working as part of an institutional—or even multi-institutional—effort. The TEI guidelines fail to provide guidance to those who have little interest in making choices, frequently have no “specialist knowledge” of the materials, but who are charged with choosing the “right” and easily shared path. I do not think that this is necessarily a problem, and particularly not when balanced with some set of recommendations for typical practice.

Let’s consider this problem from the perspective of the digital library practitioner who has more important things than figuring out complex encoding on his mind. When that person is working to try to associate page images with encoded text and tries to work within the existing TEI, he will not find a documented, approved way to get his work done. Give him a hack and make sure he knows how to use it, and he will, but he'll be suspicious about the viability and appropriateness of the approach. The next thing you know, someone else will come along with a new approach that may or may not acknowledge the TEI, and because the new approach will be clear and relevant to the digital library practitioner, he'll run with it. The encoding of the text can't be guaranteed to be consistent with the TEI, and it probably won't be. I probably *do* need to point out that there are still people out there in the Library world advocating the use of EBind instead of TEI. I don't want to be overdramatic, but that sort of decision (i.e., to use EBind) results in disastrous consequences for interchange.

Moreover, it needs to be said that because the TEI has continued to be focused primarily on the problems of the scholar, a practitioner in an institution may concluded that the TEI is not *vitally* relevant to what we do anymore. For those preservation efforts I described earlier, encoding the contents of the pages in TEI is a simple formality that could be done in just about any other way with no loss or gain of efficiencies, and as products like Meta-E move in, the page level encoding isn't even likely to remain TEI without some political wrangling. [describe Meta-E] The encoding in projects like the EEBO TCP is purposely derived from TEI, but already in its first year, it's not likely to change much. It could just as easily be dealing with mid-1990's TEI and the sort of vendor guidelines that have been used in projects like EEBO TCP and Wrights. What happens with TEI **in the future** is not likely to be of significant relevance…, unless the TEI changes.

Conclusion

If the TEI initiative were to change, to *become more relevant to institutions*, I believe it would need to change in two ways: first, it would need to promulgate a set of application guidelines; second, it would need to rise to meet a major set of challenges faced by the digital library practitioner. Where are our real challenges and, in particular, where might the TEI be relevant? The "relevant" challenges come in several areas:

1. We are frequently required to promulgate a new standard or best practice, frequently with encoding, to accommodate the relatively routine work that needs to be done. Notably, often this means dealing with new encoding issues that TEI hasn't tackled, or doesn't tackle in a sufficiently prescriptive way (e.g., compound documents).
2. We frequently focus our attention on the connections between our efforts, whether through interoperability protocols (e.g., Michigan NSF-funded distributed fulltext search project), exchange formats (e.g., METS), or syntaxes for expressing something externally (e.g., OAI and perhaps even OAIS). The TEI initiative could play a role by extending its framework.

There is, ultimately, a problem of resources. Institutions are wealthier than individuals, but we still face a need to make decisions, especially in a tighter economic climate. Institutions pay directly and indirectly to support a number of similar and sometimes overlapping efforts, from TEI to Dublin Core to OAI to METS. Institutions pay for umbrella organizations that host efforts like these. Those organizations include, for example, the Digital Library Federation, the Research Libraries Group, the Association of Research Libraries, and even the TEI. Institutions (rather than individuals) are in the best position to sustain these efforts, to pay the bills, because of the benefits the institution receives. Having interoperability and interchange between our institutions is of real and significant financial value.

Although the TEI Guidelines provide an outstanding framework for a number of textual applications, they stop short of creating a broad framework for digital libraries. Page image collections, compound documents, and mixed format repositories have become the everyday business of digital libraries throughout the world. Choices, and particularly the choices made by institutions, must be guided more formally and more actively by the TEI effort. The TEI must remain a vital part of that institutional world, and if it is to do so, it must make the Guidelines relevant to the digital library practitioner. If TEI doesn't become relevant to institutions, other organizations and standards will step in, be adopted, and TEI will have decreasing relevance while we pay the bills for those other efforts.