Issues raised for scholars in contributing all or part of a thematic research collection into library catalogs.

This grant project explored a great deal of interesting terrain, but from my viewpoint one of the most interesting issues it raised was the question of how thematic research collections fit into our overall understanding of the way digital materials are created and used.

There is an implicit or explicit rationale in the digital world, which includes two key points. First, a premise of repurposability: part of the developmental logic for digital resource collections, and an explicit part of their design, is modularization. And second, a separation of data, formatting, metadata, intellectual organization, and interface.

That's a strong design philosophy and extends not just to the practical dimension of our work but also to its larger motivations: we create digital resources in this way so that we have the liberty and the ability to separate out our different categories of information and intentions from one another and act on them distinctly, rather than all in a bundle.

Thematic research collections in particular put this design philosophy to the test, because they combine a comprehensive approach to content (which makes their content, their "raw materials" valuable from a repurposing standpoint) with a non-trivial element of scholarly shaping: they represent a perspective, an argument, a "theme". So one issue this grant required us to explore was how the use of metadata and a variety of interrelated data representations would let us capture this dimension of the thematic research collection, with the goal of supporting different kinds of digital archiving and preservation goals.

The dimension of scholarly intention and belief is of course not neatly separable from the data representation; it operates at many levels in the collection and is represented through different mechanisms. For example:

- beliefs with respect to what words or characters the scholar sees on the page (which become a transcription of content and a set of markup expressing the decisions that had to be made concerning the deciphering process)
- beliefs concerning the relation those words and characters bear to some kind of authorial (or other) reference point (which we instantiate in markup expressing things like regularization, emendation, alternative readings
- beliefs about how these words are organized into textual genres, which are also expressions of disciplinary intent with respect to the use of the text: are we treating it as a linguistic construct, as a literary work, as a material object? which gets expressed in structural markup
- beliefs about how the text is related to other texts: as a revision, a precursor, a variation, a piece of context, a member of the same genre or authorial oeuvre, and so forth: these get expressed both through metadata and by the way we assemble texts into groups
- intentions concerning how the text is to be presented to its audience, which are expressed through interface design and formalized through stylesheets, search options, sorting orders, and things of that nature.

With this level of granularization, libraries have the wherewithal to choose and control how much of this information is retained when the collection is archived, and also how much of it to display when making the collection visible.

There are a number of issues that arise for libraries acquiring thematic research collections as part of their digital repositories, and they are more open-ended questions (on which I would be very interested in hearing thoughts from the audience) than matters to be decided once and for all. First of all, is a library in fact ingesting a collection or simply objects in a collection? This, I think, is a strategic decision the library has to make at the outset and it will affect how much of the collection's infrastructure (in other words, the "delivery" and context) will be ingested along with the source materials. It's important to note that the answer will depend on how the library is positioning itself with respect to the digital collection in question: whether as a publisher (in other words, an agency competent to repurpose the materials as raw materials for future scholarship) or as a purveyor, an agency that simply conveys the finished materials to other users. And second, what responsibility does an acquiring library have to preserve or honor a scholar's intentions in creating and shaping such a collection? We need to consider several issues here:

- to what extent can these intentions be discovered or reconstructed?
- to what extent do they affect the ingestion process (e.g. the choice of materials to preserve, the kind of metadata being collected, the way in which the materials are integrated into any larger repository)?
- to what extent are they represented explicitly in a way that can itself be studied, rather than simply constituting a kind of hidden set of gotchas?

Another way of asking the same thing is, where and how are these intentions expressed? if the library wishes to archive or ingest the explicitly scholarly or interpretive dimension of the collection, is that dimension represented in what we might think of as data or metadata, or is it instantiated as well through interface choices and features? This raises the issue of where editorial information is expressed and stored: often in stylesheets, which are typically not treated as part of the "content" of the collection

This research makes a valuable contribution in a number of ways. For one thing, it makes it clearer to standards bodies that the information design that enables these kinds of distinctions is useful: in other words, we do want to be able to identify the components of the collection that represent repurposable information and those that represent the end products of scholarship. It also helps to articulate the various roles that different types of metadata can play in articulating those distinctions:

- metadata about individual digital objects (often tied closely to the object itself, as in the case of the TEI header
- metadata about processes (that document editorial ideas and activities that have inflected the representation of the digital objects)
- metadata about the structure of the collection, expressing its organization and the relationships between its component parts

- metadata about the behavior of the collection: the specific interface choices that have been made, and what they express (from the potential that is present in the collection) about specific editorial decisions.

Given these concerns, we can identify several different strategies for the ingestion of thematic research collections into digital repositories. The first we might describe as "data only": here, the library treats the collection as a set of raw materials, ignoring the loving work of the scholar but looking ahead to future scholars who may reuse the materials as the basis for a new collection with its own interface and intellectual value. This approach, implicitly or explicitly, regards scholarship itself as somewhat ephemeral or at least of less permanent value than the potential for scholarship contained in the digital materials themselves. The second approach, which we could term "connected data", takes the data as of primary importance, but also takes an interest in the connections between the digital objects and the materials they represent, and treats these connections as themselves an important part of the data to be preserved. The third approach we might describe as the "scholarly product" approach, and it takes the collection itself as of primary importance. This might be for quasi-moral reasons having to do with honoring the scholar's intentions, or it might be for practical reasons: it may be easier to simply mount the collection than to go to the additional work of providing separate access to the individual digital objects it contains. Either way, the result is to ingest and retain the collection exactly as originally designed, including the original interface and everything it implies about the scholars original conception of the collection.

This last approach retains the maximum flexibility for future access: it offers the option of treating the individual objects as of separate importance and to offer them for future redeployment, or, alternatively, to treat the collection as the only object of interest and responsibility, and offer its shape and surface as the only mode of access for the materials it contains. It also, of course, takes the maximum amount of work and opens a number of challenging issues. How can we retain the functionality of interface tools that may depend on very specific (and not future-proofed) software capabilities? How can we identify the precise nature of the scholarly and editorial commitments and intentions that are instantiated in the interface, to make it possible to retain them as the collection is migrated forwards? In other words, if we've accepted the responsibility for maintaining the collection as a scholarly artifact in itself, we've taken on a very substantial task.

And in fact this is the dimension for which there currently exist the least effective methods for documentation: it's not quite clear how a thematic research collection should express these kinds of intentions to enable them to be acted upon and preserved. Some places exist: for instance, the <projectDesc> in the TEI header (and possibly similar elements exist in METS as well? Terry?). But the problem is complicated by the fact that these intentions are often disseminated and squirreled away in a vast range of small design decisions, and instantiated in things like stylesheets, font choices, and so forth, whose significance may not be obvious except as part of the collection's overall impact. This suggests to me that there's an opportunity (and a need) for some interesting work in this area, to examine thematic research collections not from the standpoint of their development, but rather of their consumption, in order to understand the full range of ways they communicate scholarly information and intentions.