JDHCS 2009
Volume 1 Number 1

What Scholarly Editors Need to Help us Make Sense Together in the Digital Age

Peter Robinson, Institute for Textual Scholarship and Electronic Editing, University of Birmingham

I'd like to begin by imagining two scenarios. Here's the first scenario: I am sitting at a desk in, say, the Newberry Library. I am reading, as I do, the first line of the Canterbury Tales. A buzz starts up in the shelves over in the Eng Lit section: the books mutter to each other—"he's reading the Canterbury Tales!" The books ask each other, "have you got something which might help him?" Various ones answer: "I've got an edition of that bit of text;" "I've got a facsimile of a really important manuscript that has got that line he is reading;" "I've got a scholarly commentary." And by Chicago's special magic, those books pick themselves out of the shelves and fly through the air, land on the desk around me, and open themselves at just the page I want. All I have to do is turn my head slightly and there they are: exactly the edition, the manuscript page, the commentary I would like to read.

Here's the second scenario. I'm working on the text of the Canterbury Tales. Over in Trintity College, Cambridge, the excellent David McKitterick has found funding for photography of Trinity College's Tales manuscripts. In moments, those images are on my desk. I make a transcription of a few pages of the manuscript. Instantly, my transcription is on the desk of colleagues in New York, who spot some errors and correct them. Other colleagues, in Utah, run a collation of this new transcript against versions of the same text in other manuscripts, made by other scholars around the world. I see these new collations, and I remake the table of relationships we had for these manuscripts using this new data. Immediately, other scholars can see this new table of relationships; they can adjust their perceptions of certain key lines accordingly. Simultaneously, other scholars in Posnan, Poland take the transcripts and add linguistic annotation to each word in the transcripts. They then add these annotations to their databases, and other scholars can now retrieve the words they have annotated: this part of speech, that instance of this lexeme.

Of course, you know I am not describing an impossible fantasy. And in this campaign season, I can hear several of you thinking *'yes we can!'* (and some of you are ready to launch into those fine campaign chants, "Tag Baby Tag" and "T – E – I T – E – I").

Actually, you who are thinking *'yes we can'* are probably even now thinking *'actually, we can't, quite, do that yet'*. In the digital world, we can imagine a browser which sees that I am reading the first line of the Tales, and automatically forages the world to find editions, manuscripts, commentaries, and puts links to them right on the browser page. We can imagine too something like an RSS feed which says '*new images here!'* and then puts up an interface which allows me to make a transcription of the text in the image: when I click the OK button, new RSS feeds go off to the colleagues in New York, Utah and Posnan, and they get to add their bit to the scholarship.

So it is: no, we can't quite do that—yet. But yes, we could—if.

My talk then, is on the 'if'. What is it we don't yet have that would allow those scenarios to come true? Firstly, let's all offer ourselves hearty congratulations. We actually already have almost all the things we need to do this. We have computers linked to networks fast enough to shift information around the world in the wink of an Alaskan Governor's eye. (I'm sorry, I must stop these gratuitous

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

political references. But heck: scheduling this the day before a certain election, and having us meet within a mile of a certain candidate's house—that elephant just won't leave the room.) So what else do we have? We have in internet browsers the sonic screwdriver of our times: the tools which can display any data, and allow us to add to this data, and pass it to others. These are indeed miracles.

So we have mangoes and bananas, volleyball and pingpong, speeches from our sponsors and advice from Tokyo Rose, but something is missing. The two scenarios I sketch both rely on being able to carry out a single, simple operation. It's this: I want to type into a box somewhere this request: show me what you have for first line of the Canterbury Tales. I'd like to get back a list of the manuscripts which contain that first line; I'd like to get back links to digital images of the relevant pages of those manuscripts, ideally focused on just that line of text; I'd like to get back transcripts of those pages; editions, commentaries, etc. If I can just do that, all else will follow.

You can tell that I have been thinking about this for quite some time. Of course, if you spend long enough thinking, it's likely that you'll end up doing something, too. Here are a few links to articles on this at http://www.itsee.bham.ac.uk/online.htm and http://www.digitalmedievalist.org/journal/1.1/robinson/. Most recently, we have won funding for what we call the Virtual Manuscript Room: http://arts-itsee.bham.ac.uk/vmr/. On the surface, this is Yet Another Manuscript Website: we'll have hundreds of thousands of images of hundreds, eventually thousands, of manuscripts; we'll have descriptions of the manuscripts, transcripts, etc. Nothing new there. But what is new is that we are building, from the very beginning and from the ground up, the ability to answer that simple question: show me what you have for, say, verse 4 of chapter 1 of St. John's Gospel. We imagine scholars will come to our site and want to know: what pages of what manuscripts have that verse? What images do you have of those pages? What transcripts do you have of those pages?

We do this two ways: by labelling and by metadata. First: labelling. You need some way of saying: I have here an image of a manuscript page which contains the text of verse 4 of chapter 1 of St. John's Gospel. I have here a transcript of the text of that verse in that manuscript. We suggest a three part labelling scheme, which we call 'unified text identifiers'.

The first part is: labelling of texts. We say that any part of any text, down to the individual letter, can be labelled. Our scheme has two parts: the first part declares a naming authority, the second is the name itself, expressed as a hierarchical sequence of key/value pairs. For example:

"Auth=ITSEE/text=CT/": the whole of the Canterbury Tales, as defined by the naming authority ITSEE.
"Auth=ITSEE/text=CT/part=GP": the General Prologue of the Canterbury Tales, as defined by the naming authority ITSEE.
"Auth=ITSEE/text=CT/part=GP/L=1": the first line of the Canterbury Tales, as defined by the naming authority ITSEE.
"Auth=ITSEEINTF/text=GNT/book=4/chapter=1/verse=1": the first verse of the first chapter of the Gospel of John in the Greek New Testament, as defined by the naming authority ITSEEINTF.

The second part is: labelling of text sources—manuscripts, editions, print texts, any physical instance of a text. We use the same two part system, to label any part of any text source, down to a tiny space on a manuscript page. Thus:

"Auth=ITSEE-INTF/textsource=01": Codex Sinaiticus, as defined by the naming authority ITSEE-INTF
"Auth=ITSEE-INTF/textsource=01/quire=37/page=2r": the second page recto of quire 37 of Codex Sinaiticus, as defined by the naming authority ITSEE-INTF

The third part of our labelling system is: typing the resource associated with the text or text source. It is not much use saying that we have something to do with the first line of the Canterbury Tales unless we say exactly what that something is. Is it a description? An image, or a set of images? A transcription? An edition? Further, in the domain of scholarly editions, there are many editions, many transcripts, many images: and they are not alike. Thus, we use the same two-part scheme of labelling to define exactly what the resource is. Here are some examples:

"Auth=ITSEE-INTF/type=transcript/form=XML/schema=http://intf-itsee.schema" -- A transcript of a manuscript, expressed in XML and conformant to the intf-itsee schema

"Auth=ITSEE-INTF/
type=facsimile/source=microfilm/color=grey256/resolution=300dpi/form=jpg/
compression=60" -- A 256 bit grey-scale digital image of a manuscript, at 300 dpi against the original, stored in jpg form at 60% compression:

We think that this labelling scheme allows us to say exactly what we need to say about these resources. Now, we need a mechanism of telling people this. We propose to do this through metadata records, which link our descriptions of the resources using these labels to the resources themselves.

Here is a sketch of how we could do this, using Dublin Core within OAI-PMH. First, here we say that page 2 of the manuscript Mingana 10 contains the text of verses 6 and 7 of Chapter 1 of St. John's Gospel, and we point to an image of that manuscript page:

```
<dc:title>Mingana 10</dc:title>
    <msdesc:altname>10</dc:title
   <dc:identifier xsi:type="uid:textsource" >
          Auth=ITSEE/textsource=Mingana10/page=2</dc:identifier>
   <dc:identifier xsi:type="uid:text" >
      Auth=ITSEEINTF/text=GNT/book=4/chapter=1/
      verse=6</dc:identifier>
   <dc:identifier xsi:type="uid:text" >
      Auth=ITSEEINTF/text=GNT book=4/chapter=1/
      verse=7</dc:identifier>
<dc:type xsi:type="uid:type">Auth=ITSEE-INTF/
           type=facsimile/source=digitalimage/color=24bit-rgb/resolution=500dpi/
           form=jpg/compression=60</dc:type>
   <dc:identifier xsi:type="dcterms:URI" >
      http://www.vmr-itsee.bham.ac.uk/Mingana/10/2</dc:identifier>
```

Second, we say here that we have a transcript of this page of this manuscript containing the text of verses 6 and 7 of the first chapter of St. John's Gospel.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

```
<dc:title>Mingana 10</dc:title>
   <msdesc:altname>10</dc:title
   <dc:identifier xsi:type="uid:textsource" >
         Auth=ITSEE/textsource=Mingana10/page=2</dc:identifier>
   <dc:identifier xsi:type="uid:text" >
      Auth=ITSEEINTF/text=GNT/book=4/chapter=1/
      verse=6</dc:identifier>
   <dc:identifier xsi:type="uid:text" >
      Auth=ITSEEINTF/text=GNT book=4/chapter=1/
      verse=7</dc:identifier>
   <dc:type xsi:type="uid:type">Auth=ITSEE-INTF/type=pagetranscript/form=XML/
       schema=http://tei p5..</dc:type>
   <dc:identifier xsi:type="dcterms:URI" >http://url for the transcript</dc:identifier>
```

You can see the patterns here. We can group together labels so as to say: here is something about this text on this page of this manuscript; here is what that something is; and here is a URL for that something. Our hope is by feeding these through OAI and other dataharvesters, people can find the resources we label, and do what they want with them.

Some of you are probably thinking things like: Resouree Description Formats, Canonical Text Services; Xpath; OAI-ORE, semantic web, web services, etc. Without going into discussions as to how our system is different from or better than or compatible with these various systems, let me highlight a few things about what we propose.

First: what I have outlined makes no assumptions about local implementations. The resources we label might be encoded in html, xml, pdf, tiff, plain text; they might exist as separate files or scattered across databases; they might be in systems using Java, apache, python, MAMP, or anything you like. Second, you could see this as a massive exercise in standoff markup: our assertions about what the resource is, and what it contains, are separate from the resource itself. This is a fundamental distinction between what we propose and, say, Canonical Text Services or indeed most TEI-based solutions. It has immense practical benefits. For example: there are now many images of biblical manuscripts floating out there on the web: many of them held on servers which are able to host the images, but have no access to technology for indexing the images. In our scheme, a scholar could find these images and create the indexes – declaring exactly what text is on what page, even transcribing the text – and the index and transcripts would then be available to any other readers alongside the images, as if they were all on the one server within the same system.

You can see where this is leading. In this vision, we have a world-wide cottage industry drawing together scholars, readers and libraries. Various libraries and projects put up images of manuscripts and books, as and when they can. Scholars and readers see those images: some add information about the manuscripts; others go through the manuscript page by page, identifying the text on each page; others transcribe those many texts; yet others collate the text. If this sounds familiar, that's because it is familiar: this is how scholars have worked for centuries. More immediately, it is how the International Greek New Testament project has functioned for some seventy years: on the voluntary effort of many hundreds of expert and interested readers.

In this vision, the scholarly infrastructure is built from the ground up, by scholars working through web interfaces, using tools designed for their needs. At last, I have mentioned the word

infrastructure. Now, those of us who spend too much time skulking the corridors of the digital humanities know that digital infrastructure for the humanities, or 'cyberstructure', is just the hottest thing going: here is the gravy train which is going to pour all the dollars we need into the pockets of our digital humanities centres for years to come. So we are seeing projects like SEASR and BAMBOO, and funders like JISC, various e-Science programs, and the Mellon foundation, all directing themselves towards digital infrastructure. Well, I am going to dissent from this program. I think that the way many of these projects are going about building digital infrastructure for the humanities is wrong, and is a recipe for the waste of vast sums of money. Indeed, I would argue that these efforts will actually damage our work of building a really useful and sustainable infrastructure for digital scholarship, by deepening the divide between those with access to digital tools and those without, and also by increasing the cynicism about the money that goes into digital scholarship, which has already resulted in considerable disillusion in the scholarly community: the close-down of the Arts and Data Humanities Service in the UK can be seen as a direct result of this cynicism and disillusion.

There is an alternative. In the last two years, we have seen a remarkable instance of how an infrastructure can be built, from the ground up, with next to no initial resources. I'm bringing the elephant back into the room now: that infrastructure is the extraordinary Obama campaign. This morning, on my way here, I drove past the Harbor Country for Obama headquarters, on the Red Arrow Highway. Even at 6.15 am, the lights were on: and for weeks the building has been full of people, every day, from early in the morning to late at night. And here is something: all the funding for this headquarters came from the people themselves, scores of them in this one corner of Michigan. And there are thousands of such places now, all across the country.

There are several lessons here. First: the greatest forces for any infrastructure, in any community, are the enthusiasm and gifts of the people who do the work. We have in the communities interested in Chaucer, in Homer, in Dante, in the Greek New Testament, people of great enthusiasm, talent and knowledge: give them the chance to contribute, and they will. I've heard Greg Crane say that the point of the million book project is that one person cannot read a million books: that's why we need text mining, and machines to read for us. Greg is wrong. One person cannot read a million books: but a million people can, easily, in one day even. What we need is a means of allowing the million people to communicate what they find to each other.

The second lesson is: enthusiasm is not enough. As long as there have been elections, there have been throngs of individuals wanting to help. But never before have they been able to come together as they have for the Obama campaign. The difference is the internet, and the techniques the campaign has found to allow individuals to find each other: and then the resources the campaign has been able to give from the centre, so that the individuals in each centre have lists of voters, scripts for canvassing and phonebanking, literature to hand out, answers to questions. There is a central infrastructure: but its role is to enable the effort of those 'on the ground': so that their contributions are valuable, and so they feel valued.

Win or lose tomorrow: there is a consensus that the Obama ground campaign has changed politics in this country. It might be rather far-fetched to argue that it could change textual scholarship too: some of us have, for years, been arguing for the rebuilding of textual scholarship as a collaborative enterprise across the web. We who are lucky enough to work in universities have the chance, through the internet, to liberate and guide the enthusiasm of those who love the texts we are paid to work with. This seems a rather worthy aid: and, yes, we can.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

A quote;

Because SEASR is a cyberinfrastructure project, we have targeted computational humanists as our primary community, with traditional humanists as a larger, secondary community.
'Retreat report': Mellon RIT Retreat Presentation: SEASR, Present and Future, for the Princeton Retreat February 28-29, 2008; at http://seasr.org/blog/2008/02/28/making-progress-seasr-at-the-andrew-w-mellon-research-in-information-technology-retreat/
Wow.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago