Creating and Implementing an Ontology of Documents and Texts (ADHO 2018)

Introducing Textual Communities: talk by Peter Robinson at ADHO 2018, Mexico City, June 28, 2018.

First, I thank the conference organizers and all involved in arranging this conference for the opportunity to speak here today. One can derive all manner of meanings from this confluence: we are here at the historic centre of one of the great cultures of the world, and the oldest capital city in the Americas, to find our feet in a digital world driving into an unimaginable future. I'm particularly pleased to be in a session with Michael Sperberg-McQueen and Claus Huitfeldt. You will understand why, in a few minutes.

In this talk, I'm not going to show the textual communities system. Immediately following this talk, I'll be posting to various lists details of where it is and how to use it. Incidentally, don't try to find it by googling it: that will take you to the old, outdated version. The new version is functioning, it is tested, it is ready for you and others to use. I want to focus now on two aspects of what we have done. The first is intellectual, the second social.

First, the intellectual. In 2008, when I was thinking about what a cooperative and dynamic scholarly editing system might need, I went to a meeting at the University of Edinburgh on the marriage of Mercury and Philology. I outlined my ideas then of how such a system might be made, and a computer scientist there (whose name I don't recall) said to me: how you make it is not important. What is important is that you have the right model. Get that right, and everything follows. That set me to thinking. In the business of scholarly editing, there are three words which matter above all others: Document. Work. Text. What, exactly, do these three mean? How, exactly, do they relate to each other?

Like all who have lived under the shadow of the TEI, I have carved in me the formula of the Durand, Mylonas, Renear and DeRose “what is text really” article, which famously declares that text – any text – is composed of an ordered hierarchy of content objects. A tree, in fact. Dante’s *Commedia* is made up of three canticles, each divided into thirty-three or thirty-four cantos, each canto divided into around 140 lines, give or take. Right back in the earliest days of the TEI, Michael and Claus showed how the Wittgenstein Nachlass could be represented according to this model, and I followed their example in my own work on medieval English texts. This model, I now see, is exactly half right.

Over the last decade, I've been engaged in an intermittent and occasionally fervent discussion with Peter Shillingsburg, Paul Eggert, Hans-Walter Gabler and Barbara Bordalejo about the meaning of these terms: document, text, work. We are editors. We deal with texts – but also, we deal with documents. In fact, every text we ever touch exists in a document. We have no way to the text except through a document.

Of course, the authors of "What is text really", and the creators of the TEI, knew that documents and texts are intricately related. They knew that the *Commedia* exists in manuscripts and printed books. Accordingly, the TEI created a system of representing the physical characteristics of documents by the use of milestone elements: indicators of page, column and line breaks interspersed through text. Many of us found this quite adequate, and large quantities of text were encoded with this model.

We always knew there were problems with this model. Usually, this problem is expressed as “overlapping hierarchies”. The text itself is encoded in nicely ordered divisions: but the breaks between those divisions usually do not correspond with the page and other breaks. The text is one hierarchy, the document elements are scattered through that hierarchy. We can see that in this simple example:

```
<text>
  <body>
    <pb n="1r"/>
    <div n="Sample" type="entity">
      <lb/><ab n="1">This line</ab>
      <lb/>runs across several
      <lb/>lines</div>
    <pb n="1v">
      <lb/>While this
      <lb/>block runs
      <lb/>across a
      <lb/>page break.
    </div>
  </body>
</text>
```

Figure 1. A simple text with an “overlapping hierarchy”
So far, I have been talking about text as if it is—well, something we find in documents, or in what we say to each other. Usually, that is what we mean: we collapse all three terms work, text, and document so that the one word, text, might stand for all three. Over time, I have come to realize that this is not just a simplification, it is profoundly wrong.

Paul Eggert is a most hospitable person, and is fond of inviting his fellow textual scholars to come visit him in Canberra. Should you accept, he will take you for a walk in the Mount Ainsley park above his house. He will stop you in front of a tree, like this one, point to the marks, and ask you: is this a text?

![Figure 2. A “scribbly gum” tree.](image)

Well, you think: these marks could be a text. There could be a pattern. Perhaps here is a crude pictograph of an animal. It might be some relative of Thai, or some other unfamiliar language. In fact, Paul will tell you, these are marks made by a grub burrowing under the bark. Later, the bark falls away and reveals these squiggles.

As Paul Eggert puts it: this is not a text because it does not represent a human communicative act. And here is the first part of our definition of a text: it must represent an act of human communication. One may be more precise yet: we are speaking of linguistic communication. If it does not represent an act of communication, it is not a text. It is just marks on paper, or scratchings on a tree.

One may go further, and assert that indeed, each act of communication may be represented exactly as Durand and his co-authors suggested: as an ordered hierarchy of content objects. Here is where they were half right. Sometimes the act of communication may have a very simple structure: a single word or sentence. Or it may be as complex as the Bible, or the *Commedia*. But for every act of communication there is, we may detect a structure.

In our system, we call these meaningful units of communication “entities”. We do not call them “works”: works is a loaded term in textual scholarship. So, a text must be an act of communication. But that is not all a text is. An act of communication must exist in a physical form: or it does not exist at all. It must be inscribed into a manuscript, or printed on pages in a book, or recorded on a tape. Even in our thoughts it has an actual physical presence as synapses in our brain open and close. The reverse is true. Marks in a document which do not represent an act of human communication are not a text. They are just marks. Hence, this definition of a text:

A text is an instance of an act of communication inscribed in a document.
Accordingly, every text which ever existed, and ever could exist, has two fundamental characteristics. It is an instance of an act of communication, which may be represented by an ordered hierarchy of content objects. It is also inscribed in a document.

And here is where it gets interesting. A document may also be represented as an ordered hierarchy of content objects. It is made up of quires; the quires are made up of pages; the pages are made up of columns, divided into lines, divided into word spaces. So a document may be represented as a tree, as may too the act of communication.

This is why Durand and his co-authors were exactly half right. They offered a model of the act of communication as a tree. But a text is, as we define it, is not just an act of communication: it is also the document it is inscribed in. A complete representation of a text therefore presents it as two trees: one tree for the act of communication, one tree for the document.

Two trees, not one. Now let’s make things even more interesting. Consider paragraph 251 of the Parson’s Tale in the Hengwrt manuscript. Read as an act of communication, this is a single statement (or entity, as we call it):

**Figure 3.** Folios 231r and 231v of the Hengwrt manuscript, with Parson’s Tale 251

Hence, in the communication act tree: this sentence is on a single branch. But in the Hengwrt manuscript, this sentence is split across two pages, the first half appearing on the base of 231r, the second half on the top of the next page. Accordingly, this appears on two branches in the document tree.

So what, you might say. This is just an overlapping hierarchy. The text is in the same order, just differently divided. Now, look at this line:
Here is the well known phenomenon of the turned under line. In this case: the text is not in the same order in the two trees. Indeed, at this point the branches of the two trees are quite different. In the communication tree “read on this book” is in a single branch; in the document tree “on this” and “book” are separated from each other by the text “That shew of such an exercise may colour…”

If you are still doubtful: look at this example, from the Complutensian polyglot bible.
Figure 5. The Complutensian polyglot Bible.

Not only does this single document page contain multiple acts of communication: two of these, the Hebrew on the top right and the Aramaic at the bottom left, are to be read right to left. In these cases, the leaves of text in the document tree and the leaves in the communication act tree will be in reverse order.

In this model, text is a set of leaves appearing on two quite distinct trees. Each tree is completely separate from the others. On one tree, many leaves may hang in one order from a single branch. On the other tree, those leaves may be on many different branches, perhaps widely separated from each other, perhaps in a completely different order. Imagine two trees in nature that somehow have grown into each other, so that although their branches are completely distinct, yet they have the same leaves. Here is how we use this understanding for the simple text we looked at earlier:
We see that the two sentences are decomposed into seven leaves, each holding one or more words of text. These seven leaves are then distributed across two trees. In the document tree, the seven leaves are distributed across six lines in two pages. In the entity tree, the seven are distributed across two <ab> elements in one <div>.

There are many consequences of this definition of text as leaves shared by two trees. First, while each text must appear in two trees, there are many scenarios where there may be more than two trees: we might want to analyze the text rhetorically, or metaphorically, or thematically, each having its own tree. Second, the complete independence of the trees means that questions of how the trees relate to each other have no meaning. Does the communication act precede the page? Or the reverse? Do we have:

<page><communication act>…

or

<communication act><page>…

It really is irrelevant. The two exist in different dimensions. The text exists in both dimensions: but the text is all they share.

In our language, an act of communication is an entity. We represent this formally in this X-Path like syntax:

```
entity=CT:Part=GP:Line=1 is the first line of the General Prologue of the Canterbury Tales
document=Hg:folio=1r is the first folio of the Hengwrt manuscript
text=CT:Part=GP:Line=1:document=Hg:folio=1r is the text of the first line of the General Prologue as it appears on the first folio.
```

So far, the theory of documents, entities and texts behind Textual Communities. What about implementation? A fundamental requirement was that we would allow real-time editing of the two trees and the text they share. You can see the result in Textual Communities.

How did we make this happen? Well, it only took twelve years. There did not seem to be any ready-made system doing what we wanted. First, we tried to use an XML database: XML DB, now maintained by Oracle. It took four years to figure out that this would just not fly for us. In 2010,

---

Figure 6. A text with seven leaves

This text has 7 leaves:
- Document: across six lines in two pages
- Entity: across two <ab> in one <div>
we abandoned XML DB, and spent three years trying to persuade MySQL that it could do what we wanted. But our joins grew so large and complex and slow that in 2013 we abandoned this too. We experimented with SparkL and rdfs: after all, we now had a complex and rich ontology. This too could not work fast enough for what we needed. Finally, we hit upon JSON and MongoDB. And five years later, finally, here we are. For the curious, this is what the core JSON document which holds it all together looks like, for block 251 of the Parson’s Tale, as we saw it earlier:

![Figure 7. MongoDB documents representing block 251 of the Parson’s Tale](image)

Here, we see the two halves of this block. The entity tree is shown, in the blue box, in the ancestors array. Each half of the block is in the same branch of the entity tree: on this tree, they are a single text, with each half being a leaf on the branch for block 251. But the document tree, shown in beige in the docs array, is different: in the top one, the leaf is on folio 231r, in the second it is on folio 231v.

So far, the intellectually original part of the work we did. I said I would speak of the social aspect. You will find two versions of textual communities online: the “production” version and the “sandbox” version. As the name implies: the ‘production’ version offers more support, and a guarantee of data persistence, but only to a few approved projects. There are no such guarantees for the sandbox version: on the other hand, anyone can access it. So, what do you need to do to have your edition on the production version? Here is what we say:

1. Your community must be publicly visible
2. All image, transcription and collation data held in the production version MUST be available free-to-all without any restrictions (non-commercial, share-alike, derivative versions), so far as the community leader is able to grant this access. The only restriction is that anyone who takes anything from TC must acknowledge where it came from and who made it. Other people must be able to take the data, re-use it, elaborate it, alter it, republish it, make money from it: anything at all, as long as they say where the data comes from.
Behind this requirement is a view of how we should conduct ourselves as scholars in the digital age. It is the model of the open source community; what Clay Shirkey calls “design for generosity”. In this model, scholarship is not created by scholars working with all the privilege of tenure, publishing in approved journals, rewarded and regulated by learned bodies. It is created by legions: by many people reading, commenting, contributing, giving to each other. 22 years ago, a much younger Michael Sperberg-McQueen was sitting in Newark Airport after a three-day meeting on software tools for humanities scholars. Here is what he wrote about the meeting:

… the one point on which everyone seems agreed is that we need an open, extensible system, to work with texts we have not read yet, on machines that have not been built yet, performing analyses we have not invented yet. This is not a system for which we can plan the details in advance; its architecture, if we insist on calling it that, will be an emergent property of its development, not an a priori specification. We are not building a building; blueprints will get us nowhere. We are trying to cultivate a coral reef; all we can do is try to give the polyps something to attach themselves to, and watch over their growth.

I think Michael’s metaphor applies not just to the digital humanities. This is how scholarship has always proceeded. Each person adds their fragment, as best they can, to the mounting coral reef of knowledge. However, there is a catch. This can only happen if people can take from the reef what they need, make something of it, put it back, and be prepared to let someone else take what they have given. If we do this, then we can make something marvellous.

If ever I am asked “what is the value of the digital humanities” my answer is this: our great merit is that we have a culture of generosity. We give, that others may prosper; and our prosperity in turn depends on others. In thirty years of academic life, I have found nothing so destructive as the instinct of “it’s mine”: this is my data. I will control who uses it. More than anything else, the aim of Textual Communities is to offer a tool which allows scholars and readers to give to others. If people do this, we can change rather a lot.