The Background to the Textual Communities Project

The 1980s: the beginnings of Collate

The pre-history of the Textual Communities project dates back to a time before the name 'Digital Humanities' existed: in the mid 1980s, it was called 'Humanities Computing' or 'Computing in the Humanities'. There were many fewer of us, and a high proportion of the people employed full-time around the world to do what we now call 'Digital Humanities' were to be found in a single building: the Oxford University Computing Services (OUCS) on 6 Banbury Road, Oxford. The 'high proportion' was precisely two people: Lou Burnard (then known mostly for his founding and vigorous curation of the Oxford Text Archive) and Susan Hockey (then known mostly for her writings on 'Humanities Computing'). I came to OUCS in 1985, when I had started an edition of the Old Norse narrative sequence Svipdagsmál as my doctoral thesis (supervised by the legendary Ursula Dronke, alas now passed away), found myself drowning in some 40-plus manuscripts, and thought that just maybe computers might help me. I took Susan's course in Humanities Computing and met Lou.

By 1988, I was well on with my thesis, and I had been able to use a combination of various computers (first, an Amstrad PCW 256 -- anyone reading this who recognizes that name will give a sigh of nostalgia and relief; then a series of Macintoshes, beginning with a Macintosh 512, for the 512 kilobytes of memory it had; and most of all the Oxford VAX, now also deceased) to achieve a useful collation of all these manuscripts, and then a database analysis of their relationships. For this, I wrote a crude, but effective, collation program on the Oxford VAX using SNOBOL and SPITBOL (more nostalgia, more relief). These were Susan Hockey's favourite programs, and she liked what I was doing enough to suggest I write it up into two articles, published in 1989 in Literary and Linguistic Computing, on collation and textual criticism. Many elements of this work persist into Textual Communities; thus, the making of computer-readable transcripts of original sources, encoding both details of the documents (page and line breaks, etc.) and details of the work represented in the documents (poem by poem, line by line); the use of computer programs to understand this markup to carry out precise word-by-word collations of the multiple versions; the use of yet other computer-assisted methods to explore the relations among the multiple versions.

This work seemed promising enough for Susan and I to think it worth writing a grant proposal to the Leverhulme Trust, for a 'Computers and Manuscripts' project. This got funding, for three years starting in September 1989.

1989-1997 Collate and DynaText; the Canterbury Tales

Susan Hockey and I had a simple aim for this Leverhulme project: to make it possible for other people to do, on a commonly-available personal computer, what I had been able to do with Svipdagsmál using an arcane and un-repeatable combination of computers and jury-rigged programs. In 1989, before Microsoft Windows had become stable and usable, this meant only one computer: the Macintosh. Things were different then. Everything (the program, the compiler, the operating system) had to run in four megabytes of memory. All the data, and all the applications, had to fit into the 20 megabyte hard drive I had attached to the SE-30. I still have this computer and this hard drive, and every now and then I cast them a friendly glance. Our plan was that I should write a computer program, running on the Macintosh, that would allow other people to transcribe, collate, and analyze texts. This computer program became Collate.

The shift towards a system others could use caused us to look at the way we were encoding the texts we were making. We were testing the developing collation system on two core sets of texts: the manuscripts of the Old Norse Solarljóð and of Chaucer's Wife of Bath's Prologue. Vario us other people approached us with their texts at this time, notably Prue Shaw with her work on Dante's Monarchia, and we realized that we needed to develop encodings capable of coping with all these. At that time, the Text Encoding Initiative was just getting started, I became closely involved in it, and hence a participant in numerous, endless (and still-continuing) conversations with many people -- most of all Lou Burnard about encoding systems. These had a direct influence on Collate, as it took shape. I decided not to use strict TEI-SGML encoding in the transcripts we were making, and hence have Collate natively able to recognize and compare texts so encoded. There were several reasons for this, partly the persistent (and never-solved) lack of usable SGML editors and tools, partly the resistance I found when I showed scholars texts marked up with full SGML (all these brackets!). However, I did decide to make the Collate encodings 'SGML-lite', with the aim of being able to translate all materials created by Collate to SGML.

This turned out to be a fateful decision. By 1993, we had progressed so far that we now wanted to publish what we had done. SGML and the TEI were on everyone's lips. So we became committed to translating everything into TEI-SGML, and then publishing it. It turned out that there was just one publication system which could cope with this rich diet of encoded text: the DynaText application, then published by Electronic Book Technologies, and whose architect, Steve DeRose, I had also met through the TEI (Steve was also one of the creators of XML). DynaText taught us a great deal about the potential and the limitations of electronic publishing systems, and hence contributed directly to the making of Anastasia (see below). In the short-term, however, it enabled us to publish complex and large electronic books (all published by Cambridge University Press: my edition of the Wife of Bath's Prologue, Anne McDermott's edition of Johnson's Dictionary, Jim Harner's World Shakespeare Bibliography), at a time when most other systems (particularly, those on the emergent web) could manage to publish only short and simple documents.

A second fateful event dates to these years. I had chosen the Wife of Bath's Prologue as one of our testbed texts for the development of Collate largely because I love Chaucer and the Canterbury Tales, and I thought it would be pleasant to read this text several hundred times in the course of this work. I found two things in the course of working on the Wife of Bath. Firstly, the textual tradition of the Canterbury Tales is extremely problematic, and ideally suited to the range of methods which computing made available, particularly as we discovered the power of phylogenetic tools, derived from evolutionary biology, for analysis of large textual traditions (particularly, the Nature article I wrote with Chris Howe and others). Secondly, I found two other people particularly interested in what I was trying to do with the texts of the Tales: Norman Blake and Elizabeth Sologova. Norman had written widely on the Tales and the problems of its manuscripts; Elizabeth was acutely interested in Middle English manuscripts and in the intellectual challenges of their computer representation. Elizabeth and I co-wrote the seminal "Guidelines for Transcription" which guided the first years of our work on the Tales. Several statements within it are worth quoting in full:

"In the course of our work we have come to realize that no transcription of these manuscripts into computer-readable form can ever be considered "final" or "definitive." Transcription for the computer is a fundamentally interpretative activity, composed of a series of acts of
translation from one system of signs (that of the manuscript) to another (that of the computer). Accordingly, our transcripts are best judged on how useful they will be for others, rather than as an attempt to achieve a definitive transcription of these manuscripts. Will the distinctions we make in these transcripts and the information we record provide a base for work by other scholars? How might our transcripts be improved, to meet the needs of scholars now and to come?"

and:

"Any primary textual source then has its own semiotic system within it. As an embodiment of an aspect of a living natural language, it has its own complexities and ambiguities. The computer system with which one seeks to represent this text constitutes a different semiotic system, of electronic signs and distinct logical structure. The two semiotic systems are materially distinct, in that text written by hand is not the same as the text on the computer screen. They are formally distinct, in that a manuscript may contain an unlimited variety of letter forms but a computer font ordinarily will not. They are logically distinct, in that the computer transcription will attempt to resolve ambiguities present in the natural language of the primary source (e.g. the same graph being used for distinct letters; cf. the discussion of minims below): if the transcription does not do this, it will betray its principal aim of decoding of the primary source. Transcription is both decoding and encoding; the text in the computer system will not be the same as the text of the primary source. Accordingly, transcription of a primary textual source cannot be regarded as an act of substitution, but as a series of acts of translation from one semiotic system (that of the primary source) to another semiotic system (that of the computer). Like all acts of translation, it must be seen as fundamentally incomplete and fundamentally interpretative."

The key concept here, that transcription is "a series of acts of translation from one semiotic system (that of the primary source) to another semiotic system (that of the computer)" was formulated by Elizabeth Solopova. The core of this view of transcription, as the creation of a complexly-encoded text which both represents the multiple objects the transcriber sees in the text of each transcribed page and also reflects the concerns of the encoder, is fundamental to the Textual Communities project.

And, the three of us -- myself, Norman Blake and Elizabeth Solopova -- in 1992 agreed to work together on the textual tradition of the Canterbury Tales. This became the Canterbury Tales Project.

1998-2005. Anastasia; the Greek New Testament; "fluid, co-operative and distributed" editions

DynaText proved extraordinarily powerful, in publishing the Wife of Bath’s Prologue: yet there was one simple thing it could not do. It could not take a single page of our transcriptions and publish the text of that page, and that page alone, alongside the manuscript image of the page. This was the case even though we had diligently included markers for every page-break in our transcription. As every neophyte in hierarchical markup systems knows, we had hit the classic overlapping markup problem. You can encode a text according to its content by pages, or by structural elements, such as by chapter and paragraph. As soon as a paragraph split across a page break, you were bust. SGML had, in theory, a way you could include overlapping markup: the CONCUR feature, which I never did implement. XML does not have even this. We wanted to say to the program: start showing the text at the beginning of this page, go on to the end of the page, and stop. No deal. So (how hard could this be?) I set out to write a SGML/XML publication system that could do this. I also wanted to remedy some other defects in DynaText. We wanted to publish electronic books in identical form on CD-ROM and the web, from a single software base, on both PC and Macintosh computers. We wanted to show search results in KWIC (key word in context) form. We wanted to make it easier for individual scholars to publish high-quality digital editions than was then possible.

These desiderata guided the making of Anastasia. Preliminary work on Anastasia was done by John Knight, at De Montfort University, Leicester; I carried this on, with increasingly demanding parts of the programming (notably, its integration with the Apache server suite) developed by Andrew West. Anastasia was first used in the Canterbury Tales Project’s publication of Estelle Stubbs’ edition of the Hengwrt manuscript. We went on to publish some ten CD-ROMs and online materials using Anastasia, including such stars as Chris Given-Wilson’s (and others) Parliaments of Medieval England, Prue Shaw’s editions of Dante’s Monarchia and Commedia (the latter remains, to me, the most fully-realized born-digital scholarly edition yet created), and Jos Weltensberg’s Leiden Armenian Lexical Textbase.

Anastasia, alongside Collate, also became the base of my work with the Greek New Testament, first with Klaus Wachtel and others at the Institute for New Testament Textual Research in Münster, and then with David Parker and his colleagues at the University of Birmingham. I am specially grateful to Barbara Aland, then leader of the INTF, for welcoming me to the Münster fold. The Greek New Testament work, alongside my work on the transcription of the Dante materials with Prue Shaw, was critical in developing our understanding of transcription, particularly after Barbara Bordalejo joined the Canterbury Tales and other projects from 1999 on. This led to the development of a system for encoding variant texts within manuscripts, fully deployed in Shaw’s Commedia and there described by Bordalejo (Appendix C, "The Encoding System"). More significantly, this experience profoundly influenced the thinking about text in documents and in works which underlies the fundamental ‘documents, entities, texts’ architecture of Textual Communities. Further, in the course of many discussions with Klaus Wachtel, David Parker and others, we found ourselves constructing in our heads the ideal system for working collaboratively on making editions such as those produced by the Münster and Birmingham teams. There was a point to these speculations. The 5000 manuscripts (and counting) of the Greek New Testament meant that this was a task far too big for a single scholar, a single group, or even a partnership of groups. One would need many more people to contribute, and this would mean finding ways for this to happen. In Birmingham and Münster, this led to the Workspace for Collaborative Editing/Virtual Manuscript Room project. We were not alone: the same ideas inform the InterEdition project. And in Saskatchewan, they led to the Textual Communities project.

Over the same period, the web changed. ‘Web 2.0’ arrived, with its vision of the web ever-rewritten by its users. One can see how this impacted on our incubating ideas about collaborative workplaces. We wanted to make editions together; ‘we’ could be anyone; the editions we make would be everywhere, and they would change and grow by the instant as we made them. Several papers of mine from around this period contemplated these possibilities. In one of them, Where we are with electronic scholarly editions, and where we want to be, I coined the phrase ‘fluid, co-operative and distributed editions’. This is what Textual Communities is trying to make possible. (Other papers of mine discussing these issues can be read at http://usask.academia.edu/PeterRobinson.)

2005-2010 Birmingham: The Canterbury Tales project and models of collaboration
While Anastasia was a technical triumph, enabling us to publish what we wanted, just as we wanted, it did not succeed in its most profound aim: of becoming a tool usable by any scholar who might want to make a scholarly edition in digital form. By 2005 (the year I moved from De Montfort University to Birmingham), we wanted something which could create editions in real time, in collaboration with many people. Andrew West and Zeth Green both worked with me on a successor to Anastasia, with these goals in mind: this was SDPublisher, built on a very different (and much more stable and widely-accepted) combination of technologies. However, SDPublisher inherited the same basic model as Anastasia: a would-be scholar would have to install and master a rather complex piece of software. Over the same period, and particularly from 2008 when the InterEdition group became active, it became clear that web-based collaborative systems held far more promise -- an intuition completely confirmed by the starting success of Transcribe Bentham from 2010 on, and by the rise of 'crowd sourcing'. Textual Communities, as a completely web-based system, follows this model.

In this same period, we learnt a critical lesson about collaboration. In the Digital Humanities, most collaborations are painless and pleasant. Usually, the collaboration is between a subject domain expert and a technical expert: what one might call the Kings College model. Because there is a very clear division between responsibilities, attribution of credit is uncontentious and easy. However, collaborations of this kind are actually rare in the academy, where most people work most of the time with others in the same discipline. In these cases, attribution of credit in joint research can be very difficult. The long-standing preference among humanists for solitary work can be seen as a response to this -- and also means that most humanists do not have to develop methods of dealing with these difficulties. Two events in these years showed us how dangerous these issues could be. The first related to the Canterbury Tales project, which began around 1992 as an informal collaboration between Norman Blake (then at Sheffield) and myself and Elizabeth Solopova (then in Oxford). With funding from our universities, the British Academy, the Arts and Humanities Research Board and others, more people joined the project, helping transcribe manuscripts and develop materials for publication. In time, two of those involved, Michael Pidd and Estelle Stubbs, both based in Sheffield, became dissatisfied with various aspects of the project, including their sense that their work was not sufficiently acknowledged. I, and others involved in the project, thought a clear agreement had been reached with the University of Sheffield in 1998 to allow the project to further develop, and eventually publish, materials originated by the Sheffield team. Accordingly, from 1998 up to 2010 we continued to work on these, first in Leicester and then at Birmingham, with continuing encouragement from individuals in Sheffield who reassured us that we would indeed be able to publish this work. However, in 2010 the University of Sheffield informed us that they would not permit us to publish any materials originating in Sheffield, despite all the work we had done on them: in essence, transcripts of eight full manuscripts of the Tales. These have now been published by Sheffield, in 'The Norman Blake Editions of the Canterbury Tales' (http://www.chaucermss.org/; see too http://hridigital.shef.ac.uk/norman-blake-editions). To have worked so long on these materials, always in the belief that we would be able to publish them, and then find that we could not, is a searing experience. It showed, particularly, how intellectual property law could be used not to enable ground-breaking work, but as a weapon, by one person or one set of people against another.

The second event showed even more vividly how IP can be used as a weapon of intellectual destruction. I met Hans Walter Gabler, most famous as the editor of Joyce's Ulysses, first at a Society for Textual Scholarship conference in New York in 1993. He is profoundly interested in the possibilities of the digital medium, and we had many intense discussions about what we were trying to do. We saw each other many times over the years, at conference-town and other places, and became friends: Hans has the gift of friendship. Over the same period, he was involved in protracted negotiations with the Joyce estate over his desire to extend his work on the central task of his intellectual career, his remarkable 1984 edition of Ulysses. On various occasions, he sought to make a digital version of this; he sought to republish it in print; he sought just to replicate a single manuscript page. Every request was denied, with impeccable legal force, by the Joyce estate. Further, it was made very clear to him that this refusal was eternal and irreversible and he could never hope for permission to quote a single line of Joyce, ever again. This meant that he had to abandon the work of a lifetime.

Creative commons share-alike attribution and sharing

Events like these decided us. We would never work on materials where someone else could, by fiat, render all our work worthless just by refusing publication permission. Even more, we would make all the work we do as freely available as we can to others, to permit them to do whatever they wanted with it, without fear that we (or someone else) might suddenly appear and say: this is mine. It happens that the Creative Commons attribution share-alike licence matches this desire perfectly. Notice that we advocate, firmly, that the "non-commercial" restriction, beloved of many scholars, NOT be applied to this licence. We want publishers to use our material, to blend it with other resources, even to make money from it. Good luck to them.

Just declaring that something be available to everyone is worthless unless you actually do make it available. Hence, as well as apply this licence, we would also design Textual Communities in such a way that anyone can easily access all the materials we have on Textual Communities (subject to third party permissions). Specifically, we wanted to make it easy for people to extract our materials and create their own interface to them. This led to the decision to develop an Application Programmer's Interface (API) to Textual Communities. Informally, we wanted to make it possible to write a web application in less than 100 lines of code which would create an interface to everything we have. Currently, http://www.sd-editions.com/FromThePage/SimpleAlAIP.html does this in 120 lines (and we can do better).

Peter Robinson, Saskatoon, 12 November 2012- 30 April 2013, Mexico City 22 November 2013, Saskatoon May 2018.