If there is one thing that can be said about the entire literary output of the world since the invention of writing it is that literary works exist in multiple versions. Such variation may be expressed either through the existence of several copies of a work, through alterations and errors usually in a single text, or by a combination of the two. A textual feature of this degree of importance ought to be at the forefront of efforts to digitise our written cultural heritage, especially at a time when printed media are becoming less important. Until now literature has been represented digitally through systems of markup such as XML, which are ultimately derived from formal languages developed by linguists in the 1950s (Chomsky 1957; Hopcroft and Ullman 1969); but over recent years it has gradually become clear that the hierarchical structure of such languages is unable to accurately represent variation in literary text. Alan Renear (1997), for example, admits that variation is one exception that does not fit into his hierarchical model of text; likewise Vetter and McDonald (2003) conclude that markup provides ‘no entirely satisfactory method’ for representing variation in the poetry of Emily Dickinson. More general discussions of the shortcomings of hierarchical markup, including the problem of variation, have recently been made by Dino Buzzetti (2002) and Edward Vanhoutte (2004).

An alternative approach, not yet tried, is to use graphs to represent variation. Graphs were first studied in the 18th century by the Swiss mathematician Leonhard Euler, who is best remembered for his solution to the famous ‘Bridges of Königsberg’ problem (Trudeau 1993). The type of graph which most closely resembles textual variation does not appear to have yet been described by
anyone; however, it can be derived from the following example. Consider four versions of the simple sentence:

A. The quick brown fox jumps over the lazy dog.
B. The quick white rabbit jumps over the lazy dog.
C. The quick brown ferret leaps over the lazy dog.
D. The white quick rabbit playfully jumps over the dog.

Collapsing the five versions into collapsing the four versions. Such repetitions are clearly undesirable. If they were present in an electronic edition each time one copy was changed, an editor would have to check that the other copies were changed in exactly the same way. If all this redundancy is removed by collapsing the four versions wherever the text is the same, the following graph results:

![Figure 1](image1)

This is a type of ‘directed graph’, which we call a ‘textgraph’. Its key characteristics are:

a. It has one start and one end point.
b. The ‘edges’ or ‘arcs’ are labelled with a set of versions and with a fragment of text, which may be empty.
c. There are no ‘directed cycles’ or loops.
d. It is possible to follow a path from start to end for each version stored in the graph, which represents the text of that version.

In figure 1 version D contains an insertion: ‘playfully’ and a deletion ‘lazy’. These are represented in the graph as empty edges. In fact insertions and deletions are the same thing viewed from different perspectives: every deletion is an insertion in reverse and vice versa. Transpositions, as in version ‘D’ - the transposition of ‘white’ and ‘quick’ in relation to version ‘B’, can be viewed as a deletion of some text in one place and its insertion in another. All that is then needed is some way to refer back to the original text to avoid copying, e.g. by ‘pointing’ to it. This feature has been shown in figure 1 by drawing the transposed text in grey, which does not change the structure of the graph.

This model is equally applicable to variation arising from a single manuscript or from the amalgamation of multiple manuscripts of the same work. Its biggest advantage is that it can handle any amount of overlap without duplicating text. One example of a rigorous test of this model can be found in the archives of the ‘Digital Variants’ website. The poem ‘Campagna Romana’ by the modern Italian poet Valerio Magrelli exists in four drafts, the first of which is shown in figure 2.
a mechanism for recording any possible set of readings, and to leave the interpretation up to the editor.

Documents which are based on this graph structure we call ‘multi-version documents’ or mvd’s. One application of this format is the applet viewer shown in figure 3 (Schmidt, 2005). This currently allows the user to view one readable version or layer of text at a time. In reality only the differences between each layer are recorded, and the user can highlight these using red to indicate imminent deletions and blue for recent insertions. The text is also searchable through one version or all versions simultaneously. This visualisation tool is in an early stage of development and as yet it can only handle plain text. However, because it cleanly separates the content of the document (represented by the edges of the graph) from its variation (represented by the graph’s structure), the same method could also be used to record versioning information in almost any kind of document - including XML, graphical, mathematical and other formats:

This allows a multi-version document to utilise existing technology. By removing variability from a text, and effectively representing it as a separate layer, the mvd format allows technologies like XML to be used for what they were designed to do: to represent non-overlapping content. One way this could be achieved would be to edit the text in an existing editor but to modify the editor slightly so that instead of reading and writing the document directly it would read and write only one version at a time to an mvd file, as shown in figure 4.

There are a couple of possible objections to the overall technique described here. Firstly, because it is not based on markup, it is no longer practical for the user to see the contents of the document in its merged form. Secondly, existing XML technology currently utilises markup to record information about the status of individual variants. This data would have to be re-encoded as characteristics of the bits of varying text, since the document content would no longer carry any information about variation. However, the very idea of ‘variants’ embedded in the text is a structure inherited from the critical edition, which is now widely regarded as obsolescent (Ross 1996; Schreibman 2002). Through the printed medium traditional philology advanced the notion of textual ‘truth’ in its effort to restore a lost original, whereas our model is directed toward the fruition of the text as it really is. As we move forward into an age when digital text has the primary focus, some of the old ideas associated with paper-based methodologies may have to be revised or given up entirely (Fiormonte 2003).

In conclusion, the use of ‘textgraphs’ to represent variation appears to overcome the problems of redundancy and overlap inherent in current technologies, and to reduce document complexity. Thus far, a file format has been devised and has been demonstrated in a working multi-version document viewer for plain text, which is capable of representing original documents of high variability. By separating variation from content it also has the potential to leverage existing document handling technologies. This technique represents a new method of handling textual variation; it is mathematical and wholly digital in character, and unlike what it purports to replace, it is not based on the inherited structures of the printed edition.
References


Cross-Collection Searching: A Pandora’s Box or the Holy Grail?

Susan SCHREIBMAN
Gretchen GUEGUEN

Digital Collections & Research, UM Libraries

Jennifer O’BRIEN ROPER

Original Cataloging, UM Libraries

While many digital library initiatives and digital humanities centers still create collection-based projects, they are increasingly looking for ways of federating these collections, enhancing the possibilities of discovery across media and different themed-research. Facilitating access to these objects that are frequently derived from different media and formats, while belonging to different genres, and which have traditionally been described in very different ways, poses challenges that more coherently-themed collections may not.

In the last few years it has become increasingly evident to those in the digital humanities and the digital library communities, and the agencies which fund their research, that providing federated searching for the immensely rich digital resources that have been created over the past decade is a high priority. Several recent research grants speak to this issue, such as the Mellon-funded NINES: A Network Funded Initiative for Nineteenth Century Electronic Scholarship, or The Sheet Music Consortium.

While digital objects organized around a specific theme or genre typically provide opportunities for rich metadata creation, providing access to diverse collections that seem to have little in common (except that they are owned by the same institution) often poses problems in the compatibility of controlled vocabulary and metadata schema. While this problem has been noticed on much larger scales and addressed by initiatives such as z39.50 and the Open Archives Initiative’s Protocol for Metadata Harvesting, addressing the problem within a library’s or center’s own digital collections is a vital part of making such initiatives successful by leveraging cross-collection