Improving Access to Encoded Primary Texts

Terry Butler *(Terry.Butler@UAlberta.ca)*
University of Alberta

The Access Problem

An impressive amount of our literary heritage has now been put into digital editions. Much of it is encoded in XML, often using recognized standards for encoding such as the TEI. One of the primary scholarly goals behind this activity has been to increase access to the texts - by publishing them on-line, and by making the text amenable to searching. The XML tagging provides further added value for searching and display. Metadata, where it exists at all, is mostly at the collection level, or provides only a broad guide to the contents of a specific work.

Between high-level metadata access, and a direct search on the word forms of the text, there is little help for the reader. Due to the immense labour involved in creating detailed subject indexing, very few scholarly electronic texts have indexes or finding aids which would draw the reader to specific sections of the work.

To address this deficiency, a first trial has been made at automatic indexing of a substantial non-fiction work. The notebooks of Samuel Taylor Coleridge are a rich treasury of the thought of one of the 19th century's most important intellectuals. Comprised of over 6,500 individual entries (in scope ranging from a single phrase to complete essays), they are a valuable record of his thought and the active intellectual currents of the time. We have captured the text of the notebooks in electronic form, encoded with TEI. As a first step to building a coherent subject index to this material, we have generated a mapping between this material and a contemporary subject index (Roget's first edition of his celebrated *Thesaurus*).

Our strategy has been to construct connections between the conceptual categories in the *Thesaurus* and Coleridge's individual notes, based upon a weighted measure of similarity between the words of the note and the terms and sub-terms in the *Thesaurus*. Common words are weighted lightly; rarely used words heavily. Using this measure, we can connect each note to one or more thesaurus entries, which then makes the note accessible to searching through the thesaural categories. Implementing these connections through topic map technology, we have a stand-off tagging structure that relates these two encoded works but still leaves both of them unchanged, available to be delivered and shared with colleagues.

This presentation will describe the process by which we create an appropriate mapping between Coleridge's text and Roget's hierarchy, demonstrate the environment for creating and managing the stand-off tagging, and describe the utility of the resulting product.

The resulting edifice illustrates three important advantages for access to scholarly text: the index connects and relates sections of the text to larger, consistent conceptual categories; it provides access for searching that is complementary to the texts' own idiosyncratic terminology; it uses stand-off tagging to provide access without direct intervention in the electronic source text. This indexing structure, of value to researchers in its own right, is also the scaffolding upon which we will construct our subject index of the *Notebooks*, using modern terminology and accessible conceptual categories.

Background to the Project

The notebooks of Samuel Taylor Coleridge are a valuable and almost unknown resource. Much of Coleridge's work as poet, philosopher, scientist, linguist, and theologian was published only partially and fitfully in his time; the notebooks contain some of his most innovative and interesting work. They have been published in print in five large double volumes (text and notes) by Princeton University Press, with indexes to selected titles, names, and places; but there is no subject index. The intention for the series was to publish a thematic index to the whole, as volume 6. However, we argued (to the Canadian Social Sciences and Humanities Research Council, who are funding this work) that at the present time an electronic index to the work would be of much greater utility to scholars who wish to know how Coleridge's thought emerged and developed over the 40 years which these notebooks cover.

The overall goals for the project include:

- creating an accurate electronic text of the entire notebook corpus;
- creating an index and thesaurus for the notebooks which will be a start to a synthetic index to Coleridge's thought;
- providing a web-based search and discovery system which will meet the needs of scholars, making his thought on a vast variety of topics more accessible.

Bibliography

Coleridge, S.T. Ed. Kathryn Coburn. *The Collected Works of Samuel Taylor Coleridge.* Bollingen Series 75. Princeton: Princeton University Press, 1969.

Hüllen, W. *A history of Roget's thesaurus: origins, development, and design.* Oxford: Oxford University Press, 2004.

Pepper, S. *The TAO of Topic Maps.* 2001. Accessed 2005-03-15. <http://www.ontopia.net/topicmaps/materials/tao.html>

Sebastiani, F. "Machine learning in automated text categorization." *ACM Computing Surveys* 34.1 (2002): 1-47.

Thompson, H.S., and D. McKelvie. *Hyperlink semantics for standoff markup of read-only documents.* Language Technology Group, HCRC, University of Edinburgh, 1997. Accessed 2005-03-15. <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>