# Document Quality Indicators and Corpus Editions

Jeffrey A. Rydberg-Cox
University of Missouri at Kansas City
Department of English
Kansas City, MO 64110

rydbergcoxj@umkc.edu

Anne Mahoney, Gregory R. Crane
Tufts University
Perseus Project
Medford, MA 02155

{amahoney, gcrane}@perseus.tufts.edu

## ABSTRACT
Corpus editions can only be useful to scholars when users know what to expect of the texts. We argue for text quality indicators, both general and domain-specific.

## Categories and Subject Descriptors
Collaboration, Design and methodology, Communities of Use

## General Terms
Documentation, Design, Standardization, Languages, Theory

## Keywords
Editing, hypertext, corpus linguistics

## INTRODUCTION
One of the challenges faced by scholars in the humanities and digital librarians alike is the need to digitize large bodies of material relatively quickly. Humanists need their source texts in digital form if they are to study them with computational methods, while digital librarians face demands to provide electronic access to large portions of their holdings. One efficient mechanism for this sort of digital conversion involves scanning these documents, creating minimal meta-data (such as tables of contents), and providing access to the digital images. This method, however, leverages few of the advantages of an electronic environment: texts cannot be searched, documents cannot be analyzed and mined for useful information, etc. All of these methods require that documents not simply be presented as images but that they be converted to text, whether by typists or by OCR software. This conversion introduces a new set of considerations: should the texts be tagged, what DTD should be used, what kinds of information should be tagged, and so on. But this process inevitably conflicts with the initial ideal of the rapid conversion of a large body of texts into digital form.

Previously, we have suggested the ideas of a corpus editor and a corpus edition as one possible solution to the need for rapid digitization [4]. A corpus edition is a thematically coherent collection of documents whose structure and content are tagged, according to the needs of scholars, by mostly computational techniques. In our experience creating corpus editions for the Perseus Digital Library (http://www.perseus.tufts.edu), we have found that corpus-style editing allows us to produce significant collections of useful materials in relatively short periods of time. As we will discuss below, because a corpus edition relies on automatic tagging methods, some elements of these texts will not be tagged perfectly. While we believe that the level of error introduced by automatic tagging methods is acceptable, and that a large group of texts with some errors may even be preferable to a smaller collection of carefully tagged texts, this approach to editing requires the addition of an additional piece of meta-data to the digital library, a document quality indicator that allows users to see the methodology employed and tells them what level of detail and accuracy to expect from the documents in the edition.



## What Is A Corpus Edition?
A corpus edition stands in contrast to a 'clean' collection of documents with either no tagging or minimal tags preserving basic information such as page numbers or how the text was laid out on a page (i.e. *Project Gutenberg*, http://promo.net/pg/ or the *Thesaurus Linguae Graecae*, http://www.tlg.uci.edu/). A corpus edition also stands in contrast to carefully crafted electronic editions with extremely detailed tagging of a text's content, features, and context (i.e. the *Analytical Onomasticon to Ovid*, http://www.kcl.ac.uk/humanities/cch/wlm/Onomasticon/ or the *Electronic Text Corpus of Sumerian Literature*, http://www-etcsl.orient.ox.ac.uk/). The corpus editor working with a collection of texts carefully considers a minimal number of elements that should be tagged in order to make the text useful to the scholarly community (much as the designer of a hypertext system must consider how users will work with their automatically linked collections of documents [1]). A scholar working on Renaissance scientific texts must, for example, decide whether it is worthwhile to mark such formal elements of the texts as propositions, theorems, or proofs. Likewise, a person preparing an electronic edition of Shakespeare's works must decide whether to tag the original 'long s' contained in printed editions or simply to represent it as an ordinary 's'. Issues such as these exist for almost every collection of documents and the answer is not immediately obvious even to those with specialist knowledge of a field.

Decisions about what elements of a text should be tagged must always be balanced against considerations of time and scale. Corpus editions may contain dozens or hundreds of documents, representing thousands of printed pages. The corpus editor must consider not only what scholars might like to know about a text, but also which elements can practically be tagged in the large collection of documents.

Once the corpus editor has made decisions about what elements in a text should be tagged, it is then necessary to

develop scalable procedures to actually tag these elements in every text within the corpus. The requirements of scalability and relatively rapid production imply that much of this tagging must be done with computational techniques, using information extraction algorithms to identify features such as names, dates, geographic locations, street names, speakers in dramatic texts, and whatever other features are required. [6]

Because this process relies on computational techniques, we do not assume that the corpus editor will (or even ought to) proofread every tagged element in a text. Rather, the corpus editor need only proofread enough tags to ensure that the information extraction routines are working as expected. The use of computational techniques to tag a text [i.e. 3] and the development of computational editing environments for the creation of traditional editions are, of course, well known. [i.e. 2, 5] The essential difference between these projects and corpus editions lies in the belief that texts have value long before an editor considers or checks every tag in the document.

## The Need for Document Quality Indicators

One of the fundamental precepts of the corpus edition is that purely automatic markup does not introduce so much error as to obviate the advantages of the rapid conversion of a corpus to electronic form. A corpus tagged with minimal human intervention can serve as the basis for valuable tools for patrons of digital libraries and scholars in the humanities.

This precept, however, runs counter to traditional notions of scholarship in the humanities. Scholars traditionally cultivate their editions, only publishing them to a wider audience when they have approached a certain level of perfection. Scholars generally have similarly high expectations for the works that they consume. While a digital library researcher might be well pleased to produce a system that correctly identifies 95% of the geographic locations, proper names, and dates within a text, scholars trained in the tradition of detailed and careful study of texts often find the missing 5% unacceptable. When corpus-based editing is explained to users, however, many complaints disappear. Users need to know what they can expect from a text, and are often willing to accept errors if they know why they are there.

Some indicators of document quality are relatively simple and can apply to any electronic corpus. Was the text entered by hand or acquired by OCR? How thoroughly the text has been proofread? Other indicators might be relevant for only one discipline or corpus. A reader of Shakespearean texts, for example, will want to know if and how the spelling was modernized; the reader of a scientific text will want to know whether the tagged proofs were identified by hand or automatically.

Corpus editors must also document the meaning of their indicators. Does "thoroughly proofread" mean that a graduate student in the field has read the text, or that a relatively unskilled worker has checked it against a copy text? We expect that each discipline will ultimately reach a consensus about what indicators are the most important and what should be considered a high-quality text. Until this happens, corpus editors must ensure that users can find out what "good" means in a particular collection.

Document quality indicators are a form of meta-data, which must be easily available to users just as are more usual meta-data fields like the title, creator, or date. Further, this meta-data must be made available along with all the rest of the meta-data not just to end-users, but also to catalogs or 'harvesters' (in the sense of the *Open Archives Initiative*, http://www.openarchives.org).

## Laying the Groundwork for New Editions

Careful documentation of the elements and standards used in the creation of a corpus edition has another additional benefit. Corpus editions can serve as the basis of handcrafted editions at some point in the future. It will be easier for subsequent editors to begin with the automatically tagged text than to restart the process from scratch. This possibility, however, also counters traditional ideas of scholarship in the humanities. Building a new edition or commentary based on a previously marked-up text appears at first like cheating or cutting corners. Using and enhancing a corpus edition, however, is really a form of collaboration, especially when the enhanced text is returned to the digital library. Humanists will not be able to exploit the potential of corpus editions until we develop a culture that values this kind of collaboration.

## Conclusions

A corpus edition can be a useful tool for scholarship, even though its texts may contain errors. Users of these texts need to know what kinds of errors are likely and why. Each discipline will establish its own guidelines for which elements in a text should be marked, and what level of quality is acceptable. An essential part of the meta-data for each document is an indication of how it was created and how well it meets the discipline's standards for a good text. As corpus editions become more widely available, we expect further that humanists will develop new forms of collaboration based on shared electronic texts.

## REFERENCES


[1] Blunstein, J. "Methods for Evaluating the Quality of Hypertext Links" *Information Processing and Management* 33.2 (1997), 255-271.

[2] Bunker, G, Zick G. "Collaboration as a Key to Digital Library Development: High Performance Image Management at the University of Washington." D-Lib 1999. 5:3. http://www.dlib.org/dlib/march99/bunker/03bunker.html

[3] Chestnutt, David R. "The Model Editions Partnership: 'Smart Text' and Beyond" D-Lib July/August 1997 http://www.dlib.org/dlib/july97/07chesnutt.html.

[4] Crane, G. and Rydberg-Cox J. "New technologies and new roles: the need for corpus editors". in *Proceedings of the 5 ACM Conference on Digital Libraries*, 2001, ACM Press, 252-253.

[5] Lecolinet, E. Likforman-Sulem, L Robert, L. Role F. and Lebrave, J-L.; "An Integrated Reading and Editing Environment for Scholarly Research on Literary Works and their Handwritten Sources" *Proceedings of the Third ACM Conference on Digital Libraries*, 1998, ACM Press, 144-151.

[6] Rydberg-Cox, J. Chavez, R., Smith, D., Mahoney, A. and Crane, G. "Knowledge Management in the Perseus Digital Library" *Ariadne*, 25 (2000), http://www.ariadne.ac.uk/issue25/rydberg-cox/.