Recovering Information from Stored Text

Frank Wm. Tompa

Department of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1

fwtompa@uwaterloo.ca

Much scholarship relies on access to documents and produces new documents. Widespread adoption of computing environments has resulted in scholarship being expressed with and through computers. Scholarship therefore increasingly depends on electronic document repositories and the growth of so-called digital libraries. We are moving towards libraries as envisioned by Vannevar Bush in 1945.[1]

As in physical libraries, the documents to be housed in scolarly collections include historical documents, literary works, reference texts, and government publications. Even more apparent in computer-readable form are collections of business documents (from annual reports and customer literature to procedure manuals and internal communications) and linguistic corpora (collections of spoken and written communication assembled to reflect the uses of language). The so-called gray literature, including technical reports, personal communications, and online help information, also constitutes a growing text resource.

These simple observations lead us to recognize diverse goals for storing texts.

- Digital libraries must preserve textual materials. Howard Webber, formerly with Houghton-Mifflin and now with Digital Equipment, warned attendees at an early conference sponsored by Waterloo's Centre for the New OED and Text Research that printers' widespread practice to destroy typesetting tapes was akin to cultural annihilation. Text data in computer-manageable form must remain accessible indefinitely for unforeseen users and unpredictable purposes.
- Digital libraries must provide access to textual materials. At the digital libraries workshop held recently in New Jersey, Elizabeth Liddy said this best: "This is the fundamental challenge of the information age — not to simply add to the deluge of information, but to satisfy people's information needs by enabling them to find and use the specific information they need, when they need it, in a form they can use."[2]

---

[1] "As We May Think," *Atlantic Monthly*, July 1945, pp. 101-108.

[2] Elizabeth D. Liddy, Michael B. Eisenberg, Charles R. McClure, Kim Mills, Susan Mernit, and James D. Luckett, "The Intelligent Digital Library," distributed at *1994 Workshop on Digital Libraries: Current Issues*, Rutgers University, May 19-20, 1994. Available from liddy@mailbox.syr.edu

- Digital libraries should also become a forum for social interactions. Computers should not only support individual access to stored materials, but they should also provide communications channels for sharing information with others.
- Electronic document repositories must support document creation and enhancement. New materials will be created from scratch, assembled completely from existing materials, or extended explicitly from parts of other documents. Facilities must be provided to combine text access, scholarly collaboration, document evolution (with edit trails), and text deposit. Furthermore, electronic document management must not restrict nor inhibit scholarly expression through limitations in static and dynamic document models. For example, critical editions of texts must be supported, including rich mechanisms for critical apparatus and for maintaining variants.

There are many facets of texts that are manifest in paper-based documents. Chief among these are the information content (the concepts embodied in the text), linguistic expression (the formulation in natural language), document structure (the logical organization in paragraphs, sections, footnotes, and so forth), document layout (the physical organization, including pagination and text placement), and graphical rendering (the surface image of the text, including typography). In converting the medium to computer-based storage, document designers need to choose which of these to maintain, which to note, and which to ignore.

For example, one can choose to maintain a text as a bitmap or as a sequence of characters. The former explicitly maintains the graphical rendering of a document; it is equally suited to all printed images (not merely text) and provides a simple mechanism for browsing documents in familiar form. The latter explicitly maintains the linguistic expression; it supports text search and extraction, and with appropriate additional notation, can encode document structure and layout.[3]

Neither of these two choices are well-suited to representing the numerous relationships that occur within texts and among texts. For example, there are explicit references and cross-references such as footnotes, bibliographic references, annotations, references to figures and tables or to other sections, and references to variant texts; there are assemblies of references in the form of tables of contents and back-of-the-book indexes; and there are implicit references through allusions and by related content. In response to this need, document models typically support hypertexts, through which documents and document fragments are linked together in an arbitrarily complex web. Just as authors and publishers have adopted conventions for signaling references and cross-references in paper-base media, conventions must be developed for supporting links in electronic documents.[4]

To accommodate the wide variety of character-based representations used for text, a standard descriptive language has been introduced. The Standard Generalized Markup Language (SGML) defines how classes of documents are to be described (through a grammar capturing the

---

[3] For an empirical comparison of the usability of print, bitmap, and character sequence, see Dennis E. Egan, Michael E. Lesk, R. Daniel Ketchum, Carol C. Lochbaum, Joel R. Remde, Michael Littman, and Thomas K. Landauer, "Hypertext for the Electronic Library? CORE Sample Results," *Proceedings of the Third ACM Conference on Hypertext*, San Antonio, December 15-18, 1991, pp. 299-312.

[4] Frank Halasz and Mayer Schwartz, "The Dexter Hypertext Reference Model," *Communications of the ACM 37*, 2 (February 1994), pp. 30-39.

document type definition, or DTD) and how an individual document's structure is to be marked (using a concrete syntax). For example, the following declarations describe a bibliographic citation as an interleaving of authors, references, dates, and free text, where a reference is (optionally) the word "in" followed by the name of a work, optionally followed by an edition, and all other fields are arbitrary character sequences.

```
<! ELEMENT    citation                               --     ( author | ref | date | free_text )+ >
<! ELEMENT    ref                                    OO     ( "in"? , work , edition? ) >
<! ELEMENT    (author , date , work , edition)       --     PCDATA >
<! ELEMENT    free_text                              OO     PCDATA >

<! ATTLIST    work
              form  (monograph, serial, article)            monograph >
```

The last two lines indicates that elements tagged as "work" may be marked by an attribute (named "form") taking a value to indicate whether the work is a monograph, serial, or article, with the default being "monograph." With these declarations, a citation may be marked as follows:

```
<citation>
  <author>Frank Wm. Tompa</author>,
  <author>G. Elizabeth Blake</author>,
  <author>Darrell R. Raymond</author>,
  <work form=article>Hypertext by link-resolving components</work>
  <ref>in <work>Hypertext '93 Proceedings</work></ref>,
  Seattle,
  <date>November 14-18, 1993</date>,
  pp. 118-130.
</citation>
```

In this example, each structural element in the text is tagged using SGML's characteristic reference concrete syntax consisting of matched, angle-bracketed tags. The declarations of elements "ref" and "free_text" indicate that corresponding tags are optional ("OO" in place of "--" used in the DTD); the attribute-value pair "form=monograph" is omitted from the second work element.

SGML thus provides a method to represent a complex document as a sequence of characters recording the linguistic expression (as a sequence of words) augmented by markup for encoding document structure, layout, or other chosen features of the text. To enable text interchange among computer systems having various representations for individual characters, SGML also supports the declaration of character sets (such as ASCII or Unicode) and provides a mechanism for representing additional glyphs (e.g., `†` to represent the character †). Notably, SGML does *not* define a particular encoding or markup, but leaves this to individual SGML applications.

One recently defined application of SGML has been developed by the Text Encoding Initiative (TEI), sponsored by the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Literary and Linguistic Computing (ALLC).[5]

[5] *Guidelines for Electronic Text Encoding and Interchange*, C. M. Sperberg-McQueen and Lou Burnard (eds.), ACH/ACL/ALLC Text Encoding Initiative, Chicago and Oxford, May 1994.

> The impetus for the project came from the humanities computing community, which sought a common encoding scheme for complex textual structures in order to reduce the diversity of existing encoding practices, simplify processing by machine, and encourage the sharing of electronic texts.... [It] documents various base tag sets: these include specialized tags for prose, for verse, for drama and other performance materials, for spoken materials, as well as for letters and memoranda, printed dictionaries, and terminological data.... [Additional] topics covered include a variety of approaches to the analysis and interpretation of texts, and include representations for hypertextual links and other non-hierarchic structures, as well as specialized tags for the encoding of critical editions and language corpora.

The TEI Guidelines provide comprehensive conventions for representing documents and thus establish a basis for scholarly communications.

Beyond text storage, however, communications support also requires document manipulation. Two major text management tools have emerged from Waterloo's participation in computerizing the *Oxford English Dictionary*: PAT™ is a flexible, efficient full-text search engine capable of locating all occurrences of arbitrary strings in very large texts, and LECTOR™ is a flexible, interactive text display system capable of rendering text with embedded markup according to dynamically selected style sheets and responding to users' interactive "pokes" at the text.[6] This software has since been extended and enhanced, and both tools are now commercially available from Open Text Corporation, a University spin-off company located in Waterloo.

Many problems must be solved before Bush's dream can be realized. Electronic document collections must grow significantly, requiring voluminous data entry and extensive markup effort. Software must be extended to provide more functionality in recovering information implicitly stored in texts — for example, neither information content nor graphical rendering are easily extracted from TEI-encoded documents; the former requires sophisticated natural language processing and the latter depends in preserving bitmap images. Document database software must be developed to ensure convenient access, both immediacy and permanence must be supported. Procedures for providing document security must also be developed, so that only authorized users can read and update documents and document authenticity can be assured. As a related, but non-technical development, document owners and government agencies must establish guidelines and procedures to support commercial use as well as fair use of their intellectual property when it is embodied in computer-based media. As data storage capacity continues to increase (current technology provides 6.5 gigabytes on an optical disk; projections include 1 terabyte on a credit card in the near future) and computer communications technology attempts to keep pace, recovering information from stored texts remains a challenging problem.

Presented at the Annual General Meeting of the Royal Society of Canada, Calgary, June 11, 1994.

---

[6] Frank Wm. Tompa, "An Overview of Waterloo's Database Software for the *OED*," *Proceedings of the Symposium on Historical Dictionary Databases and Data Retrieval Requirements* (Toronto, October 1991), in *CCH (Centre for Computing in the Humanities) Working Papers 2* (1992) pp.123-143.