The KWIC and the Dead: A Lesson in Computing History

Neville Holmes, University of Tasmania

A culture can be defined as "the sum total of ways of living built up by a group of human beings, which is transmitted from one generation to another" (*The Macquarie Dictionary*, The Macquarie Library Pty. Ltd., Macquarie University, Australia, 1997). In this sense, a profession is a culture, or at least a subculture, defined by its purpose and responsibilities. To fulfill that purpose better, the profession should cultivate its history.

The computing profession's purpose is to promote, for the greater community's benefit, the use of formal representations of facts or ideas and of machines and processes for the storage and transformation of such representations. Thus, our profession's history originated in the culture of the printers and scribes who promoted the use of the written languages from which our present binary representations developed.

George Santayana wrote that those who cannot remember the past are condemned to repeat it. We can derive two warnings from his observation. First, what we think to be innovations will often be mere repetitions. Second, our profession can develop faster and better through cumulative innovation, building on its past instead of ignoring it.

REPETITIVE INNOVATION

About six months ago, British Telecom discovered it had previously been granted worldwide patents with claims that could be considered to cover the use of hyperlinks. Of these, BT cited US patent 4,873,662—which grants a monopoly until 2006—to seek license fees from Internet service providers in the US.

Hyperlinks, included in the body of electromagnetically encoded documents, refer to other electromagnetically encoded documents. Many commentators observed publicly that hyperlinks very closely resemble the endnotes and footnotes long used by scholars. BT's ambitions seem to have softly and suddenly vanished away thanks to this publicity.

Yet even if we disagree that the electromagnetic encoding of index markings follows obviously from the tradition of footnotes and endnotes, or from the allied tradition of embedded scholastic citation, more recent history gives a more specific precedent.

In his very famous article, "As We May Think" (*Atlantic Monthly*, July 1945, http://www.theatlantic.com/unbound/flashbks/computer/bushf.com), Vannevar Bush described in some detail a conceptual machine called the *memex*. According to Bush, an essential feature of the memex, a personal library machine, was that it could perform "associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another."

The memex plainly foreshadowed BT's patent, which presumably would not have been granted if the "inventor" or patent examiner had read Bush's article. That article was neither obscurely published nor ignored. It inspired many developments, such as Theodore Nelson's early Project Xanadu (http://www.xanadu.net/), which in turn inspired Tim Berners-Lee's World Wide Web.

CUMULATIVE INNOVATION

We cannot use the hyperlinks in memex documents, or in the items stored on the Web, without a mechanism for moving from item to item along a trail of links, and they are of only limited use without a means of sharing trails and their items.

Bush envisaged making microfilm updateable in situ through dry photography, which would allow not only the insertion of links at will, but also of marginal notes and comments. The memex would encode each item with an identifier that photoelectric cells could swiftly locate. Two items to be linked would be brought up in separate "viewing positions"—1945's version of Windows. The user could then cross-copy the links from the item identifiers and later use the memex to track along either link between the items. Links became visible for selection as the name of a trail, allowing an item to have more than one link. Following a memex trail strongly resembled the way browser software jumps from URL to URL today.

Although much cruder, Bush's item and trail sharing—in which the user "sets a reproducer in action, photographs the whole trail out, and passes it to his friend

Continued on page 142

Ideas published decades ago may suggest solutions to some of today's most pressing technological and social challenges.

Continued from page 144

```
profess  231:26, 389:16
professes  276:8
profession  25:14, 99:33, 154:14,
   181:24, 304:40, 389:36, 437:5,
   480:29, 489:23
professionally  178:3
professions        489:20, 546:40
```

Figure 1. Sample index in traditional concordance format, showing only page and item numbers for each word indexed.

for insertion in his own memex, there to be linked into the more general trail"—clearly anticipated the Internet hypertext transfer protocol's support for browser software using URLs.

The misnamed browser

Thus, browsers do with modern techniques what Bush's memex was to have done with microforms. But browsers, however useful, are poorly named: They do not allow true free-association browsing, only page display and link-and-trail tracking.

A drawback to such trail following is that it strictly confines users to existing trails. Using a browser is like shuttling from clearing to clearing in a forest. At any clearing you can only choose from established paths marked by signposts that bear cryptically brief descriptions. Chancing upon an interesting trail remains difficult. Although Bush foresaw "a new profession of trailblazers," a memex user had to rely on a "code book" to find a trail by name.

Link trails hold the Web together. Browsers allow us to wander along them, viewing items as we go. But the Web's trails remain weaker than the trails Bush envisioned. Formed from inherently one-way links, the Web's trails are unlabeled, typically short and highly branched, often very localized, and usually adventitious and unreliable.

Searching for Xanadu

Limitations like these motivate the continuation of Project Xanadu. Started in 1960, Xanadu and various subsequent "virtual library" projects seek to follow Bush's lead. Such projects strive to build a disciplined nook in the Web, even though most Web material remains woefully undisciplined.

Search engines help us pluck trails from among the Web's general chaos. Playing the role of Bush's code book, search engines usually provide three services: index building, hierarchical classification, and index querying. Commercial search engines index client organization Web sites, while sites such as Yahoo offer general Web users an index to the entire Web—an ever more ambitious task. By the end of last year, for example, the Google Web site claimed to have indexed 1.3 billion Web pages.

These information retrieval services have evolved from decades-old innovations. The index building programs, now often called Web crawlers, build inverted files, pioneered largely by Gerard Salton and first used for text retrieval in the 1960s. Hans Peter Luhn introduced automatic hierarchical classification, while index inquiry sprang from Luhn's work on selective dissemination of information, started in the 1950s.

Profess: politics like ours p.	231:26
p. and call themselves	389:16
Professes to flatter	276:8
Profession: Adam's p.	437:5
charmed me from my p.	480:29
contrary to their p.	389:36
debtor to his p.	25:14
head of the literary p.	181:24
most ancient p.	304:40
ornament to her p.	99:33
panted for a liberal p.	154:14
parentage is a very important p.	489:23
Professionally he declines and falls	178:3
Professions: all p. are conspiracies	489:20
p. which are full	546:40

Figure 2. Excerpt from* The Oxford Dictionary of Quotations *index, showing an expanded format that gives readers more context for each entry.

Modern search engines—the products of cumulative innovation—have applied techniques from the past 40 years to Web-based text. As such, these engines suffer from inherent limitations: Those who use their results can either work their way through a hierarchical tree of topics with link lists at the end of branches or can submit search requests that return a list of links. Link lists serve many purposes well, but generally cannot be used for deep research or even to satisfy everyday curiosity. Further, users cannot browse searched text itself, unless they download and display the entire Web page that contains the text.

HISTORICAL INNOVATION

Many researchers still frequent traditional libraries because browsing works best when it takes you down trails you didn't expect or leads to trails others might never have found. Such trails come into being when unanticipated associations occur among ideas through the act of browsing itself.

Query results from search engines provide a limited kind of text browsing, somewhat like using a book index or traditional concordance. In such an index, keywords appear in lexical sequence with a page number or other locator. Figure 1 shows the style of a traditional concordance, which provides only a page and item number for each entry.

Lists such as this one lack local context, a shortcoming that has been overcome by expanding the concordance, as shown in Figure 2, drawn from *The Oxford Dictionary of Quotations of 1941* index. Such indexes resemble the link result list of a Web search query and serve many purposes. They work even better with more context than shown here.

In his work on automatic indexing by computer in the 1950s and later, Luhn proposed two kinds of indexes: keyword out of context (KWOC), which resembled Figure 2's example but gave more context, and keyword in context (KWIC), which provided a rearrangement of KWOC.

Christians. // All who	profess and call themselves	389:16
less. // They politics like ours	profess, / The greater prey upon the	231:26
of a dedication is flattery: it	professes to flatter. // The known sty-	276:8
-makers; they hold up Adam's	**profession**. // There is no ancient ge-	437:5
that are contrary to their	**profession**. // Those things	389:36
man a debtor to his	**profession**. // I hold every	25:14
-buttons—-I panted for a liberal	**profession**. // My father was an emi-	154:14
the head of the literary	**profession**. // Your Majesty is	**181:24**
An ornament to her	**profession**.	99:33
almost charmed me from my	**profession**, by persuading me to it.	480:29
Parentage is a very important	**profession**; but no test of fitness for	489:23
a member of the most ancient	**profession** in the world. // Lalun is	304:40
friend he drops into poetry. //	Professionally he declines and falls,	178:3
against the laity. // All	professions are conspiracies	489:20
Doing-good, that is one of the	professions which are full. Moreover	546:40

Figure 3. Possible hyperindex application that would enable true browsing, built around Hans Peter Luhn's keyword-in-context automated indexing techniques.

The KWIC index, also known as the permuted title index when used for indexing article titles, proved the more enduring. Some search engines still offer a KWIC formatting option for their results, and Lexis has even been able to register KWIC as a trademark for its implementation of this option, implying an understandable historical ignorance on the trademarks registrar's part and a naïve cupidity on the registrant's part.

Yet even with extra local context, the KWIC index offers only a marginal improvement over the more familiar search-result link list. The innovation that would make the index much more effective for true browsing must do for KWIC what hyperlinks did for footnotes: make their text active in what might be called a *hyperindex*. In a hyperindex the individual words can be used, for example, to augment or refine the index being viewed or to view another section. I believe that a KWIC hyperindex, such as that hinted at in Figure 3—while an obvious development of the prior art and thus not patentable—has yet to be adopted anywhere. For further description of how such a hyperindex might work, see http://www.comp.utas.edu.au/users/nholmes/hyperindex/.

Why have KWIC hyperindexes not been implemented or, if implemented, why have they not been widely adopted? Why do word processor programs not offer a KWIC hyperindex option for a document in progress so that authors can check their style as they go? Quite simply, the computing profession at large remains largely unaware of the history of automatic information retrieval. This oversight has important repercussions not only for our profession, but for the communities we serve.

The KWIC hyperindex offers a much easier interface for executing any but the most basic queries. Ordinary users do not readily grasp query expressions, so the availability of a more usable query interface that also provides a true browsing capability would undoubtedly boost the value of textual material stored informally on the Web.

Many groups are striving to build organized, Web-based collections of educational and cultural text. Yet we face the very real danger that the Web's imminent commercialization will swamp these collections, which could rapidly make obsolete or altogether block such content. Effective hyperindexes could enhance the usefulness of these altruistic and culturally invaluable collections, making it more difficult for commercial and government interests to overwhelm the Web with e-commerce and video-on-demand content. ✸

***Neville Holmes** is an honorary research associate and a lecturer under contract at the University of Tasmania's School of Computing. Contact him at neville.holmes@utas.edu.au.*

Circulation: *Computer* (ISSN 0018-9162) is published monthly by the IEEE Computer Society. **IEEE Headquarters**, Three Park Avenue, 17th Floor, New York, NY 10016-5997; **IEEE Computer Society Publications Office**, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720-1314; voice +1 714 821 8380; fax +1 714 821 4010; **IEEE Computer Society Headquarters**, 1730 Massachusetts Ave. NW, Washington, DC 20036-1903. IEEE Computer Society membership includes $14 for a subscription to *Computer* magazine ($14 for students). Nonmember subscription rate available upon request. Single-copy prices: members $10.00; nonmembers $20.00. This magazine is also available in microfiche form.

Postmaster: Send undelivered copies and address changes to *Computer*, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Canadian GST #125634188. Canada Post Publications Mail (Canadian Distribution) Agreement Number 0487910. Printed in USA.

Editorial: Unless otherwise stated, bylined articles, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *Computer* does not necessarily constitute endorsement by the IEEE or the Computer Society. All submissions are subject to editing for style, clarity, and space.

COMPUTER
Innovative technology for computer professionals