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/flashbk/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 updatable in situ through dry photogra-
for insertion in his own memex, there to be linked into the more
general trail”—clearly anticipated the Internet hypertext trans-
fer 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, how-
ever useful, are poorly named: They do not allow true free-associa-
tion 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 profes-
sion 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 wan-
der 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, typi-
cally short and highly branched, often very localized, and usu-
ally adventitious and unreliable.

Searching for Xanadu

Limitations like these motivate the continuation of Project
Xanadu. Started in 1960, Xanadu and various subsequent “vir-
tual 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, hierar-
chical classification, and index querying. Commercial search
generates 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 ex-
ample, 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 classifica-
tion, while index inquiry sprang from Luhn’s work on selec-
tive dissemination of information, started in the 1950s.

Modern search engines—the products of cumulative inno-
vation—have applied techniques from the past 40 years to Web-

**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.
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 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.

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.

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