ALEXANDER STREET PRESS

Articles

April 2001

A different direction for electronic publishers – how indexing can increase functionality

Stephen Rhind-Tutt, April 2001

In the past three years, there has been an increasing trend toward extremely large databases. Companies like Gale, Bell & Howell, EBSCO and Wilson now have merged what used to be independent collections of journals into "mega files" with as many as six thousand titles. Internet services such as Google, AltaVista, and Lycos now index as many 1.4 billion pages, and are moving to add further large collections of texts, such as newsgroup archives.

At the same, time there's been a move away from precision searching, along with increased pressure to create a single user interface to all of this material. Worried that their users are unable to do Boolean searching or understand how to navigate multiple search screens, librarians are requesting ever-simpler interfaces. Vendors have responded. Almost all of the systems listed above open their services with a single entry box for searching. The user is simply required to key in a word. Behind-the-scenes technologies such as natural language mapping and relevance ranking serve to improve the search results.

The combination of ever-larger files and ever-simpler queries lead us to systems intended for the average user and the average query. Such systems miss the fundamental purpose of certain queries - for example when scholars are looking for rare, "off the beaten track" information that has not been found before. Such systems provide little incentive to publishers or to librarians to incorporate new information and new ways of finding it.

It is not my intent to disparage these kinds of systems - several of which proved useful in writing this article. It is to suggest that such systems are only part of the picture. We need also to look at systems that provide richer research and analysis of data, where the goal is to do more than simply retrieve articles in response to keyword searches.

This article describes an alternative approach to electronic databases. Instead of creating a system that attempts to answer general questions from many different users on many subjects, this approach focuses on enabling a particular group of users to answer in-depth questions in a specific discipline. Instead of relying on automation to reduce the need for human intervention, this approach requires substantial intellectual effort. The examples I use are restricted to the humanities and the social sciences, but they could equally well be used in other disciplines.

Background

About a year ago, Pat Lawry, Eileen Lawrence and I left Chadwyck-Healey following its acquisition by Bell and Howell. Our experience sets had been varied. Eileen knew a wide range of library systems from her time with what was then Ameritech. As a librarian and a professional indexer, Pat had a rich understanding of how indexing could be used to improve a system. I'd had experience managing the InfoTrac product line, as well as seven years experience with SilverPlatter Information. In this last role I had seen how the indexing and controlled vocabularies used by files like MEDLINE could enrich systems and make them considerably more efficient.

As we surveyed the electronic publishing scene, several things were apparent:

- The vast majority of systems were designed to provide "pretty good" retrieval of articles corresponding to a wide range of subjects. These systems excelled for specific keyword searches. They did less well in orchestrating the results of queries. And very few had the ability to combine concepts (contrasted with combining words) to restrict searches.

- The increasing size of the databases meant that traditional notions of selection and quality were driven by the editors of the original works, rather than by the needs of the aggregated database. With the exception of purposefully built encyclopedias such as Grove's Dictionary of Music, most systems relied on their size to overcome deficiencies of omission and expected the user to select which article to read from a list of candidates.

- The absence of precision in search results precluded certain kinds of secondary analysis. For example, a question such as, "Which author was the first person to use the word abortion?" could not be asked unless the system was capable of generating a precise list of authors.

- Very few database publishers were investing in indexing. Instead they were relying on keyword searching, in some cases supplemented by basic date, title, author and (sometimes) subject fields. This puzzled us, because the larger the file the more important indexing becomes.

- Indexing tended to be source-work-centric rather than user-centric. In other words, the system would consider a web page or a work or a chapter as the answer to all queries. Instead of listing authors, events, characters, or specific items, the system always assumed that the user wanted a document. Questions such as, "Give me all battles in which more than 200 people were killed," or "List all events that correspond to these

criteria," could not be asked.

- Several full-text databases required the user to ask questions, rather than presenting the user with choices. A number had a browse list of subjects, but many users wanted a browse list of authors, journal titles, places, media types, and more. This tied directly to the deficiencies of indexing. With multiple index fields and controlled vocabularies, it is relatively easy to present the user with the complete contents of the database in an organized, tabular form. However, if the contents have been indexed only by subject, then there is no easy way to display, say, the geographic locations contained in the database.

- The rigorous construction so evident in bibliographic journal databases like MEDLINE was not being applied to most new databases, especially in the case of electronic texts. This presented an opportunity, especially since the development of processing power and software tools over the past few years has enabled large, multi-field, relational databases to be built more cheaply than before.

In all of our research there was one system that was conspicuously different - The Internet Movie Database (www.imdb.com). This system enables searching for characters, for movies, for actors/actresses, rather than just for movies. It has browse tables, enabling you to view the contents of the database by country, by date of release, by language, and more. It allows the user to ask sophisticated questions such as "Give me all movies that star Dustin Hoffman and Tom Cruise". This may seem like a simple question, but it requires the system to have standardized descriptions of actors, and to understand the relationship between actor and film.

Developing a product

In July 2000 we decided to begin a new publishing company, Alexander Street Press. Our goal was to develop databases that would address the opportunities above.

We began by asking scholars and researchers what kinds of questions they wanted a database to answer. Our first product was decided quickly - it was to be North American Women's Letters and Diaries. We chose this partly in response to market demand, but also because the letters and diaries lent themselves to extensive indexing. Although these writings contained valuable information, it was extremely hard to locate what you wanted to find.

One of our early decisions was on the issue of what were to be the basic elements of the database. These are the items that would be returned to users in response to queries. We decided on three files - one for authors, one for sources, and one for documents. We defined a document as a month of diary entries or a letter.

I'm going to dwell on this for a moment, because it's an excellent example of the kind of decision that is critical in creating a different kind of application. Rather than choosing a month of diary entries as the basic unit, we could have chosen a chapter of a book or a single day's entry in a diary.

The former - a whole chapter - would have dramatically reduced the value of the database, because there would have been no way for us to make chapters of information correspond to dates. This in turn would have meant that users could not perform searches such as, "Give me everything written in May 1835," or any other questions having to do with dates.

The latter - a day's writing as the result unit - would have meant that even the simplest search would yield excessively high numbers of hits. The average size of a daily diary entry is a few lines, so the user would be forced to wade through hundreds of tiny entries.

Our choice of a month of diary entries or a letter as the basic element for the document file allowed us to preserve the integrity of dates. It also allowed us to separate materials written after the initial entry, so that searches would yield items written contemporaneously.

Field Specifications

The initial specification called for some 40 fields to be available for searching each letter and diary. Although we didn't anticipate having quite this number of fields, it quickly became apparent that scholars wanted them. Some of the fields were obvious requirements. Others were added only after we understood how users would want to search the database. For example, the ability to restrict document searches to letters sent to men was of interest to sociologists exploring the distinctions of gender.

We noticed early on that having so many fields enabled us to calculate meta-data that was useful in itself. For example, using the Date Written and Date of Birth fields, we were able to calculate the Age at Time of Writing field. This provides a rich avenue for new questions to be asked. For example, the ability to search by Marital Status and combine that with the Age of Writer allows the user to see the attitudes expressed by young women when they were first married.

The data creation was done using a team of seven professional indexers. Pat Lawry and Laura Gosling, librarians experienced in the development of thesauri, created controlled vocabularies, standardizing names, places, subjects and more.

The net result of this work is a relational database with three major files and more than eighty fields. Some forty of these fields are searchable by the end user. The other fields serve to enhance the display of materials. Each file within the database - author, source and document - is independently searchable. The user is able to ask questions such as, "Give me all publications in the database from the Pennsylvania Historical Society" and retrieve a list of source works. One can also, for example, ask for a list of all authors in the database born in Pittsburgh who were married and had more than 5 children. Searching is also possible by document ("Give me everything written on April 5th, 1892").

A different direction?

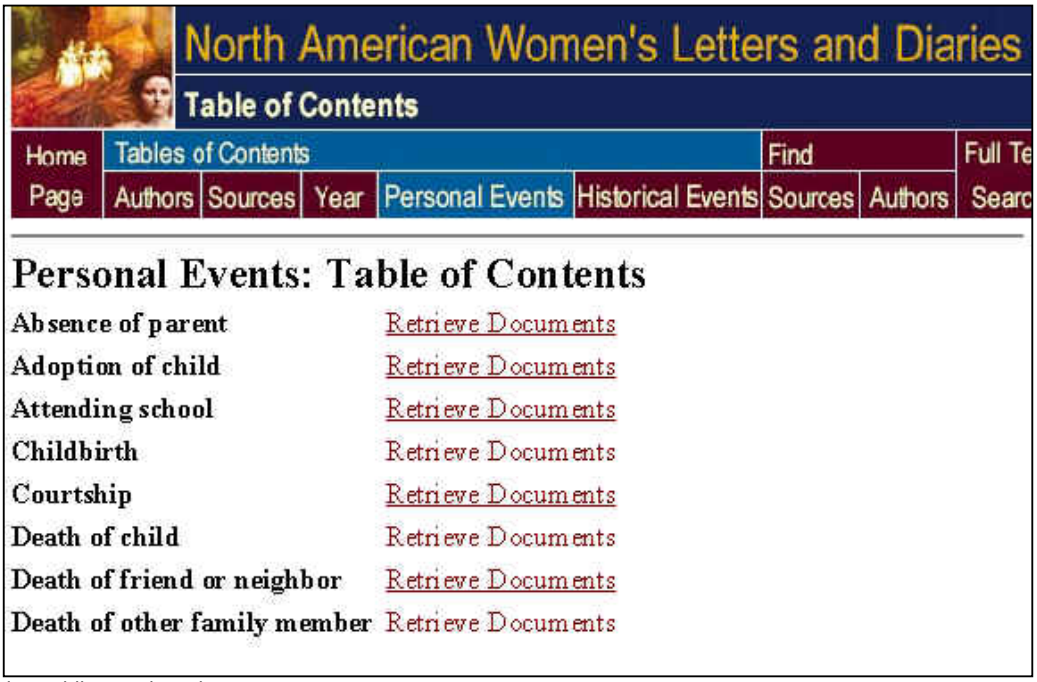
It's fair to ask whether this kind of database is really so different from the multitude of files already being used by libraries. I believe the answer is yes. Each of the four products we are producing - North American Women's Letters and Diaries, Civil War Letters and Diaries, Early Encounters in North America in North America, and American Film Scripts Online - has more fields by an order of magnitude and tailored definitions of what constitutes a document, and each product enables a different level of query than just about anything else available.

From a technical standpoint, each of these databases is fully relational, containing multiple, separate, interlinked files. For example, in the case of the Civil War database, there is an Events file that enables the user to ask questions such as, "Give me a list of all battles with more than 300 casualties." One can also search to find documents pertaining to those battles. In American Film Scripts Online there is a character file, enabling sociologists to search and retrieve a list of all African American characters who played in films from 1950-1960, for example.

From a patron standpoint, the level of query moves research significantly further than otherwise would be possible. The intellectual effort that we expend in categorizing the material enables users to examine hypotheses much more quickly than before. Using the available fields, the user is able instantaneously to extract the materials for answering the following questions:

- Does a woman who has many children have a different attitude about the death of her child as compared with a woman who has few children?
- Were the encounters between the Jesuits and the Huron more violent than those between the Franciscans and the Huron?
- How did attitudes toward slavery among women on plantations evolve in the years after Reconstruction?
- What metaphors are common in first-time encounters between European explorers and Native Americans?

The system's indexing allows scholars to conduct their research more quickly and at a greater depth than ever before. This level of indexing also enables a layperson to understand and use the database in a richer fashion, because it shows the contents of the database in ways that would otherwise remain hidden.



(actual list continues)

For example, the Table of Contents listed above - one of five available TOCs - enables the user to view the contents of the database organized by personal events (key events in the life span of a woman), inspiring users to view materials in ways that they might not have thought of independently.

Where next?

The cost of indexing at this level is extremely high. One might ask whether the utility we provide is worth all of the investment. My view is that it depends on the application you're trying to create.

The products we've created reflect the natural role of librarians. Some library users rarely talk to the librarian, while top-level research has always relied on librarian involvement. Highly indexed databases allow for the same scenario - some users will do a keyword search, find a few documents, and be satisfied, while top-level researchers will benefit from the increased utility. Our databases offer both simple and advanced search screens, to support both groups.

We already have large, expensive systems capable of giving us the journal articles we need. What we lack are systems that provide scholars and laypeople new ways of exploring, analyzing, and discovering information. For the novice, such new systems can provide easy ways to find what they're looking for, and for the imaginative researcher they can lead to unique, new understandings that were not possible in print alone.