Academic Projects

University of Chicago
Creating a "convivial" database interface

In French, the phrase user friendly translates as convivial. The project on American and French Research on the Treasury of the French Language (ARTFL) at the University of Chicago is using the NeXTstep development environment to make a French-language database, Le Tresor de la Langue Francais (Treasury of the French Language), or TLF, more convivial and more useful to a broad range of scholars.

The TLF dates back to 1957 when the French government initiated an effort to create a new, comprehensive French-language dictionary, modeled in part on the Oxford English Dictionary. The Centre National de la Recherche Scientifique (CNRS) took on the task of compiling the dictionary. The goal was not to look at individual words as they developed historically, but rather to create the first in a series of synchronic dictionaries.

Robert Morrissey, director of ARTFL, describes the project: "A synchronic dictionary focuses on words in the context of usage in a single period. You might have a dictionary for the seventeenth century and one for the eighteenth century. In this case, the goal of the French government was to create a dictionary of modern [nineteenth and twentieth centuries] usage with extraordinary breadth and depth.

"Another goal was to make use of computers. It took 32 people
10 years to enter all the literature selected for inclusion. Examples for this dictionary are being drawn from this body of text." The result, according to Morrissey, is the largest natural language database of its kind.

Today, the ARTFL database extends from the works of the nineteenth and twentieth centuries to eighteenth and even seventeenth century works and it contains approximately 2,000 texts totaling 150 million words. Works range from biographies to essays, novels, economics treatises, plays, poetry, diaries, and scientific journals.

Broadening access to ARTFL
In 1981, ARTFL began offering American researchers on-line access to ARTFL through a server located in Chicago. Originally, the database resided on an IBM mainframe and was accessed via WATS lines. However, this equipment allowed only one part of the database to be searched at a time, and search requests had to be submitted one day in advance. ARTFL also created a consortium of institutions whose researchers wanted access to the database. For a feeÐ$500 per year for research institutions, $250 for other institutionsÐparticipants receive full, unlimited on-line access to the database. Currently, 30 schools are enrolled, including Yale, Princeton, Harvard, Stanford, University of California at Berkeley, Duke, Dartmouth, and University of Michigan. Today, researchers log in to the TLF via the Internet, a worldwide network.

Creating a more useful database
To increase access to the database, programmers at the University of Chicago created a database server that allows complex searches. The new database server allows proximity searchesÐsearches that let users look for occurrences of keywords within a defined range. For instance, users could search for "liberty" and then search for every occurrence of "equality" within 100 words of "liberty." The new server can also provide a variety of statistical information concerning words and text. The server is implemented using innovative functional programming techniques, which allows users to examine preliminary results before the entire query is computed.

Taking the NeXTstep
Using Interface Builder, Nicholas Burke is creating a NeXT interface. Traditionally, searching a large body of text involves repeated selection. First, you select all the works of a certain time period. From these works, you select all the works of a certain author. From these works you select all the works containing a certain word. Eventually, through repeated selection, you narrow the body of text to a few works. Because you made your selection by entering sequences of text commands like "Select Author Flaubert," it was difficult to remember what you already selected. Narrowing your selection and completing the search was often slow and confusing.

The NeXT interface speeds process by presenting the works in the database in a hierarchical browser similar to the File Viewer in the Workspace Manager.In the top level of the browser, the user might specify

several time periods, 1700s, 1800s, and 1900s, for example. Selecting 1700s causes the next column in the browser to display the names of all the authors of works from the 1700s. Selecting an author's name causes the next column in the browser to display the titles of all the works by the selected author. Selecting a title displays the text of the work. At any time in the selection process, you can select several items in the browser and select entire collections of works for keywords.

You could choose to search "liberty" in all the works of the 1700s. Or, if you were interested only in the works of a particular author from the 1700s, you would select 1700 in the first column of the browser and the author's name in the second column. Now when you search for "liberty," you will be searching only the selected author's works and only the author's works from the 1700s.

With a browser, you can adjust your selection, and you can always see what category you selected. To narrow your selection first by author and then by time period rather than the opposite, you can set up a browser with author's names in the first column and the periods in the second.

The interface also provides advanced features such as query tools and an easy way to build word lists. One query tool is a wildcard tool that lets users enter codes to represent several words. For example, a user could enter the code "liber*" which means to search for all words that begin with liber. Word lists let users search the database for more than a single keyword. A researcher searching for liberty, for example, might create a word list containing liberty, equality, and fraternity. Now, the researcher can quickly search for occurrences of all three words, thereby searching for the eighteenth century concept of liberty rather than the word liberty.

Moving to CD-ROM

ARTFL is creating a CD-ROM-based version of the database designed to reside at individual institutions because many users expressed desire to control the database locally. The CD-ROM will offer access through a NeXT system to that portion of the database not covered by copyrights. Requests too complex for the local machine or that involve copyrighted materials will still be handled via the Internet.

In the context of this dual, distributed approach, Morrissey continues, "The NeXT environment is a perfect fit: a UNIX machine that humanists can feel comfortable with. It networks easily, comes with a CD-ROM reader, is reasonably priced, and offers great interface design tools. You could say that the NeXT is both

user friendly and developer friendlyÐideal for our distribution strategy."

The project is nearing completion. A working prototype of the interface was shown at an annual meeting between CNRS and CILS in June 1991, and ARTFL expects to release the CD-ROM and NeXT interface at the end of this year.

For more information, contact:
Mark Olsen
Assistant Director, ARTFL
mark@gide.uchicago.edu