Encoding Independent Database Applications in Tamil

Ravindran K. Paul
Micro Mart, Malaysia
www.thunaivan.com

Introduction

The past 20 years have seen great strides being made in the development of Tamil computing. From the first DOS based word processors to the current software with dictionaries and spell checkers, great improvements have been made.

Despite all these developments the bulk of Tamil software development is still related to text processing. Whether it is a spell checker, a search engine or an OCR, it is still primarily concerned with text processing. The result of all these programs are text documents.

In the commercial world, the situation is quite different. The primary task is in processing data, especially numbers. The output is numerical, graphical or otherwise. This is the area that Tamil has still been unable to penetrate.

Databases

The heart of most commercial applications is the database. The Internet itself is a vast repository of data and would not be able to function if not for the many databases working together to keep it together. Similarly the "bread and butter" software used in most organisations are dominated by database applications.

There has been some development in one aspect of database applications and that is search functionality. But, these systems are relatively easy to design, as basically the task of searching is to find a unique sequence of text within a larger sequence of data. There are several simple search algorithms and search engines that easily accomplish this task.

The problem that arises when designing truly database oriented applications is sorting.
This is where Tamil has serious problems. The structure of the language makes this task quite difficult. The current use of 40 to100+ characters to represent the 200+ Tamil characters make sorting difficult. This applies for both 8 bit and Unicode encoding. In current systems, a single Tamil "alphabet" is represented by anything from one to three characters. Worse still modifiers can fall before and after consonants makes simple sorting especially difficult.

Most commercial applications naturally use sorting. This includes dictionaries, membership systems, library systems and the like. There are already several Tamil applications that have been created to perform these functions using search alone. The difficulty is in including the added convenience of sorted databases.

Encoding

An additional issue in Tamil is the presence of 4 official encoding, namely TSCII, TAB, TAM and Unicode not taking into account the various popular non-standard encoding. Current popular wisdom states that the solution to this problem is to abandon all encoding except Unicode. This way, all resources can be concentrated on creating database functionality for Tamil without duplicating efforts to create separate functions for each encoding.

But, is this the right solution? The primary problem here is the choice of Unicode. Any programmer knows that the manner in which Unicode is stored makes processing text incredibly difficult. Add to it the fact that Unicode text can be stored in more than one way and this becomes even more difficult. The use of several bytes per character is an added challenge.

The use of 8 bit encoding for database applications is definitely easier. There are already a few applications that have data sorting capabilities using 8 bit formats. The primary reason would be the ease with which the data can be analysed and manipulated, both visually and physically. Data can be easily manipulated. The data can be easily viewed and displayed using almost any programming tool without requiring any special software or modules.

The Solution

The solution to this problem would be to take the exact opposite approach. The sorting capability should be separated from the encoding. This way database functionality can be utilised independent of the encoding. This has the added advantage of allowing current data and Tamil enabling applications to be immediately employed without having to upgrade hardware or operating systems.

It is important to understand that Tamil encoding as it is represented in its current form cannot be sorted. Since none of the current encoding can be sorted we can conclude that sorting is independent of encoding. In other words, irrespective of whether we use TAM, TAB, TSCII or Unicode, the data still requires an intermediate process to sort. In 8-bit encoding irrespective of which encoding is used, the process will still be the same. In other words, with a little extra effort, Tamil databases can function independent of encoding.

Unicode on the other hand would require more complex handling, though not insurmountable. On the other hand if Tamil Unicode had 300+ slots as proposed by some, sorting becomes a trivial exercise. Alternatively, rearranging the representation of Tamil characters within the current 64 character block would also solve this problem.

The purpose of this paper and presentation is to show that such applications are possible. The presentation will show some encoding independent applications running with several popular Tamil encoding in use today.