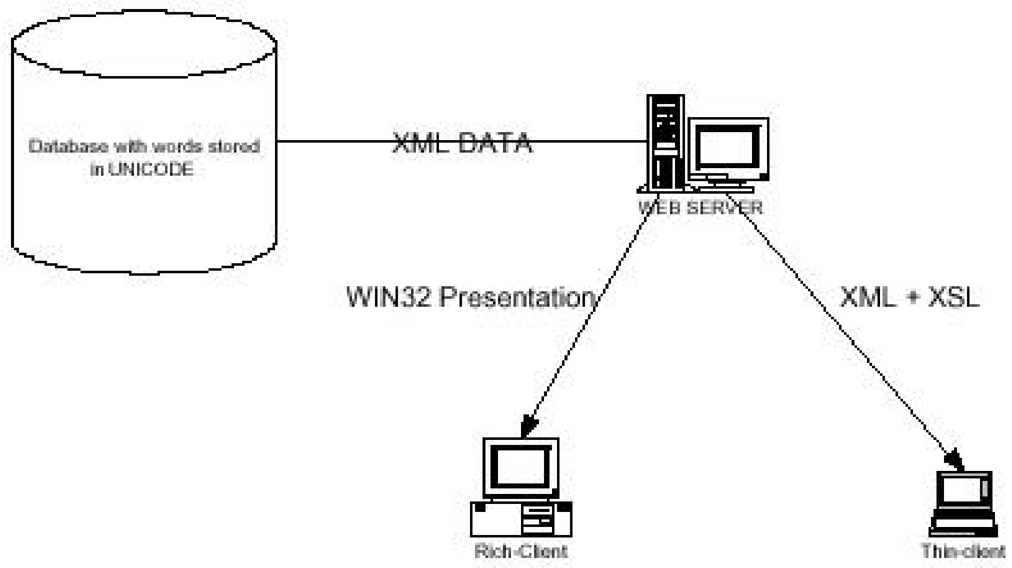
XML and Tamil text processing with Unicode

T.N.C.Venkata Rangan & Dr. Satheesh
Vishwak Associates, Chennai
Email: < Venkat@vishwak.com >

Objective:

To demonstrate seamless movement of language-independent data using XML and Unicode. The data, both in English and Tamil, will be stored in a UNICODE compliant database like SQL Server 2000, in Unicode format and will be retrieved in XML. The XML data will then be displayed in different clients. Depending on the client, we will use different XSL or other presentation techniques.



What the demo will show?

The demonstration will include a fully- working model of a English-Tamil dictionary. The user accessing the dictionary can query for a word in English and could get its meaning in English and in Tamil. The user could be using either a Rich-Client or a Thin-Client. The demo would prove the usability of UNICODE in Tamil web applications.

What is Unicode?

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. The Unicode Standard has been adopted by such industry leaders as Apple, HP, IBM, JustSystem, Microsoft, Oracle, SAP, Sun, Sybase, Unisys and many others. Unicode is required by modern standards such as

XML, ECMAScript (JavaScript), LDAP, WML, etc. It is supported in many operating systems, all modern browsers, and many other products.

What is XML?

XML is the `Extensible Markup Language'. XML is not a single, predefined markup language: it's a meta language -- a language for describing other languages -- which lets you design your own markup. XML is designed to make it easy and straightforward to use data on the Web: easy to define document types, easy to author and manage documents, and easy to transmit and share them across the Web.

HTML was all about presentation and was confined to web-browers only, where as XML clearly separates the data and the presentation. Using XML we can have one source of data and present in different browsers or different clients using multiple style sheet (called XSL). XML is fully UNICODE compliant; this enables handling of non- latin characters including Indian Language easy and safe over the wire.