Data Converters
A Case Study

Overview
This document presents a case study on the "Data Converters" suite that Webdunia has developed and maintained since 2003

Situation

For years since computers entered the Indian PSUs, Central and State government departments, publishing houses as well as the next door DTP shops, majority of the data fed into them has been in proprietary ASCII based TTF fonts. There were numerous ASCII fonts available for each language and each had its own code point mapping scheme not following any standard order even if the fonts in question were developed by the same vendor. This resulted in creation of huge databases, books, and data repositories highly incompatible with each other, neither searchable nor fit for sorting. With the internet revolution on the doorstep, this data wasn't fit enough for coming up as search results. It could not be shared across applications and platforms. This data was useful for printing only.

Solution

The worldwide encoding standard "Unicode" provides unique code points to every alphabet of the world and Indian languages are not an exception. Unicode has so far (Ver. 5.1.0) provided the standard values to alphabets, symbols and other characters in all of the following Indian scripts:

1. Devanagari (Hindi, Marathi, Nepali and their dialects as well as Sanskrit)
2. Bengali (Assamese and Bengali, other minority languages, of northeastern India.)
3. Gurumukhi (Punjabi)
4. Gujarati
5. Oriya
6. Tamil
7. Telugu
8. Kannada
9. Malayalam

The solution to this problem was a data conversion tool which could take as data created in any of the prevalent fonts and convert to Unicode accurately, without losing its semantics, requiring very small or no effort of getting it manually proofread.

Webdunia Data converter suite being developed and maintained since 2003 fulfills these requirements besides facilitating conversion from other standards such as ISCII.

Product

Webdunia Data Converters is a suite of components that provides API for inter-conversions of Indian language between following formats:

- ISCII
- Unicode
- UTF-8
- Phonetic

TTF (ASCII font based converter) - Encodings of all popular fonts of major font vendors in India for all of the following languages

- Hindi
- Marathi
- Gujarati
- Tamil
- Telugu
- Kannada
- Malayalam
- Bengali
- Assamese
- Oriya
- Punjabi

Webdunia Data converters are available as COM based components which can be directly integrated with Windows based desktop applications and usable as Web services for internet/intranet based applications on any platform.

Webdunia has developed a data conversion tool TBIL using the data converters at its backend. This tool is made available for download from Bhasha India. For in-house usage as well as for several clients, the data converters are integrated with various forms of Content Management and data flow systems.

A demonstration is available free for use at:

http://utilities.webdunia.com/dataconversion.php

Business Benefits

1. These components can be used by desktop applications /web applications for converting legacy data in fonts and other standards to Unicode.
2. These components can be used by search engines to convert crawled data from font based sites and save them as Unicode.
3. As font based input is still prevalent due to lack of Unicode/Indic support in various desktop applications these components can be used by such apps to store/export data as Unicode.