The Challenge of International Data Quality and Unicode

Vish Vishwanath Ph.D
International Market Director, Firstlogic, Inc.
vish.vishwanath@firstlogic.com

Executive Summary/Abstract: There is nothing as delightful as experiencing unrelenting growth of your customer base. As this happens, one of the immediate challenges you encounter is that of consolidating customer information into a repository that provides a clean enterprise-wide view of your customer base. Imagine the insights that lurk in such a database. There is only one hitch to unlocking these insights. Your massive repository has records from numerous countries in numerous writing systems, that embody locale-specific conventions from San Francisco to Shanghai. This presentation will explore the nuances of creating the information quality underpinnings that will help you unlock.

Overview

- The case for global data integration and Unicode
- IT challenges of global data integration
- Unicode
  - What is it?
  - What it’s not
- Information quality challenges of global data
- Global data integration strategy

The Case….

- Globalization of markets
- Geographic market expansion
  - Emerging “middleclass”
- Increase customer intimacy
  - Serving existing customers better
- Lowering the cost of global IT infrastructure
  - Consolidation of regionalized information processing centers
- Unicode enables global information integration

The Challenges….

- Language and Writing Systems
- Spoken languages share a writing system
  - English, Spanish, French, Italian, German, etc. are Latin
- Spoken languages use multiple writing systems e.g. Japan
  - Kanji (Chinese chars used in Japan)
  - Kana (Hiragana - & Katakana - )
  - Latin for proper names

Your Customer Database

<table>
<thead>
<tr>
<th>Locale</th>
<th>Writing Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>Latin</td>
</tr>
<tr>
<td>Taiwan</td>
<td>Latin</td>
</tr>
<tr>
<td>Japan</td>
<td>Latin, Hiragana, Katakana, Kanji</td>
</tr>
<tr>
<td>Korea</td>
<td>Latin, Hanja (JaME)</td>
</tr>
<tr>
<td>Vietnam</td>
<td>Latin, chu Nom, chu Han</td>
</tr>
</tbody>
</table>
The Challenges ....

Writing system basics

- Alphabetic systems – "a character represents a unique sound"
  - E.g. Latin, Greek, Cyrillic and Armenian
- Syllabic systems – "a character represents a whole syllable"
  - E.g. Devanagari and Thai
- Ideographic systems – "characters represent things or an idea"
  - E.g. Kanji
- Many characters are shared across writing systems

Sample Language/Writing System combinations...

- Adygei Cyrillic
- Afrikaans Latin
- Ainu Katakana, Latin Japan
- Aizo Cyrillic
- Albanian Latin [2]
- Altai Cyrillic
- Amharic Ethiopic, Ethiopia
- Amo Latin Nigeria
- Arabic Arabic
- Armenian Armenian, Syriac [3]
- Assamese Bengali, Bangladesh, India
- Assyrian (modern) Syriac
- Avar Cyrillic
- Awadhi Devanagari India, Nepal
- Aymara Latin Peru
- Azari Cyrillic
- Azerbaijani Arabic, Cyrillic, Latin
- Badaga Tamil India
- Baghdi Devanagari India, Nepal
- Tahtian Latin
- Tajik Arabic [3], Latin, Cyrillic (→ Latin) (also Tadzhik)
- Tamaights Tifinagh [1], Latin
- Tamil Tamil
- Tai Cyrillic
- Tatar Cyrillic
- Telugu Telugu
- Thai Thai
- Tibetan Tibetan
- Tigre Ethiopic, Eritrea, Sudan
- Tdashki (see Cherokee)
- Tulu Tamil
- Tvar Cyrillic
- Tuzar Syriac (see Syriac)
- Udekhe Cyrillic
- Udmurt Cyrillic

Computer systems use character sets

- Logical groups of characters used for a specific purpose
- Language(s), geographic, writing system, computer architecture, application
  - ASCII – American Standard Code for Information Interchange
  - ISO 8859 family
  - Far East – JIS X 0201 and Shift JIS
  - China, Vietnamese, Indian.....
  - Each having a different collection of charters and ordering

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

Unicode defines a universally accepted standard

- Offers significant cost savings over the use of legacy character sets
  - Enables a single software product to be used across multiple platforms, languages and countries without re-engineering
  - Allows data to be transported through many different systems without corruption
- Defined and maintained by the Unicode Consortium ("ISO like")
  - Unicode relatively new (Version 1.0 – Released in Oct. 1991)
  - Latest Version 3.2
- Adopted and supported by industry leaders

Unicode – Encoding

- Common character groupings
  - Does not infer a specific writing system
  - Writing systems often share characters
Unicode is not…. 

- A formatting/visual standard
- Internationalization
  - Language/writing system combinations
  - No automatic language translation
  - No localized interfaces
  - No automatic numeric, date, time, currency conversions
- Writing system translation
- Not necessarily “double byte”

Global Customer Data Nuances

Japanese Address

〒100-0004 東京都 千代田区 大手町 2 - 3 - 1

Post Code

Address:
- Prefecture
- City (Shi)
- Ward (Ku)
- District (Chome)
- Block (Ban)
- Building (Go)

Global Information Quality

- Customer data is culturally specific
- Data errors and correction are context specific
  - Phonetics
  - Misspellings
  - Typos
- Incorrect data mappings/transforms
- Cross writing system matching
- Regional data models

Variations of firm dash marker

Block data marker

“no” marker

Global Customer Data Nuances …more than meets the eye
So, What Will It Take?

- Strong technology underpinnings
  - Ability to read, write, and process Unicode data
  - Flexible technology implementation that supports adding new countries expeditiously, and supports changing conventions
  - Ability to integrate effortlessly into your enterprise data architecture
  - Integrated globally – processed locally

- Global experience and attitude
  - Locale-specific knowledge
  - Relationships/partnerships around the globe

- World-wide referential data
  - Breadth of coverage (# of countries covered)
  - Depth of coverage (data to the address level)
  - Types of coverage (address, geo, etc.)

Summary …

Success is dependent on mission-critical, data-driven initiatives (CRM, ERP, DW, BI, etc.).

Information quality foundation will influence success or failure.

Global enterprise models will require a global approach to managing information quality.

Through 2005, over 50% of data warehouse and CRM deployments will suffer limited acceptance, if not outright failure, due to lack of attention to data quality issues (0.8 probability) … Gartner Inc., T. Friedman