

8th International Conference on Information Quality, 2003

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 .

1

8th International Conference on Information Quality, 2003

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

2

8th International Conference on Information Quality, 2003

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



3

8th International Conference on Information Quality, 2003

Your Customer Database



123 Main Street
La Crosse, WI
54601

Kirkcudbrightshire
252001, m. Kaiti-1, uyt. Xpenarrrm, 22

290 King Street
Castle Douglas
Kirkcudbrightshire
DG71HA
UK

1 Avenue du Canada
91947 Les Ulis
France

180-0002 東京都武蔵野市吉祥寺東町2丁目25-4

4

8th International Conference on Information Quality, 2003

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 愛 (Chinese chars used in Japan)
 - Kana (Hiragana - あ & Katakana - ア)
 - Latin for proper names

5

8th International Conference on Information Quality, 2003

The Challenges

- Japan – 4 writing systems – often intermingled in same sentence – Hiragana and Katakana are native to Japan – Hiragana is phonetic representation - Katakana is for words borrowed from English, French Japan (encodingu, Ringu, etc.) - Kanji character set borrowed from China
- Hanzi (China and Taiwan) – Kanji (Japan) – Hanja (Korea) – chu_Han(Vietnam) use character sets originating in China – tens of thousands of characters
 - Readings differ based on language and context
 - Readings may differ but meanings may be same
 - A Japanese person may *understand* a Chinese representation but may not be able to say it right.
- Input methods via keyboard – prone to error

Locale	Writing Systems		
China	Latin	Zhuyin	Hanzi(s)
Taiwan	Latin	Zhuyin	Hanzi(t)
Japan	Latin	Hiragana Katakana	Kanji
Korea	Latin	Hangul (Jamo)	Hanja
Vietnam	Latin	chu Nom	chu Han

6

8th International Conference on Information Quality, 2003

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

7

8th International Conference on Information Quality, 2003

Sample Language/Writing System combinations...

- Adygei Cyrillic
- Afrikaans Latin
- Ainu Katakana, Latin Japan
- Aisor 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
- Azeri Cyrillic, Latin
- Azerbaijani Arabic, Cyrillic, Latin
- Badaga Tamil India
- Bagheli Devanagari India, Nepal
- Tahitian Latin
- Tajik Arabic [3], Latin, Cyrillic (→ Latin) (aka Tadzhik)
- Tamazight Tifinagh [1], Latin
- Tamil Tamil
- Tat Cyrillic
- Tatar Cyrillic
- Telugu Telugu
- Thai Thai
- Tibetan Tibetan
- Tigre Ethiopic Eritrea, Sudan
- Tsalagi (see Cherokee)
- Tulu Kannada India
- Turkish Arabic [3], Latin
- Turkmen Arabic [3], Latin, Cyrillic (→ Latin)
- Tuva Cyrillic
- Turoyo Syriac (see Syriac)
- Udekhe Cyrillic
- Udmurt Cyrillic

8

8th International Conference on Information Quality, 2003

The Challenges

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 characters and ordering

9

8th International Conference on Information Quality, 2003

Unicode

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.

10

8th International Conference on Information Quality, 2003

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

11

8th International Conference on Information Quality, 2003

Unicode – Encoding

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

12

8th International Conference on Information Quality, 2003

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"

13

8th International Conference on Information Quality, 2003

Global Customer Data Nuances Japanese Address

〒100-0004 東京都 千代田区 大手町 2-3-1

Tokyo-to (prefecture) Chiyoda-ku Ward Ote-machi District Block-sub-block-Building

Post Code

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

14

8th International Conference on Information Quality, 2003

Global Customer Data Nuances Japanese Address

2-3-1 Ote-machi
Chiyoda-ku, TOKYO 100-0004

15

8th International Conference on Information Quality, 2003

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

16

8th International Conference on Information Quality, 2003

Global Customer Data Nuances ...more than meets the eye

Variations of firm dash marker

OCRM Org Number	Org Name	Alias Name	Organization Address 1- street address (line 1)	Organization Address 1- building Name (line 2)	Organization Address 1- Postal Code	Organization Address 1- city Name
1+40890	(株)C I J	シーアイジエイ	西區平沼1-2-2 4	横浜N Tビル5F	220-0023	横浜市
2000483	(株)C I J	シーアイジエイ	西區平沼1-2-2 4	横浜N Tビル	220-0023	横浜市
2000483	(株)C I J	シーアイジエイ	西區平沼1-2-2 4	横浜N Tビル	220-0023	横浜市

Block data marker

2000483	(株)C I J	シーアイジエイ	西區平沼1-2-2 4	横浜N Tビル	220-0023	横浜市
2000479	(株)C I J	シーアイジエイ	西區平沼1丁目2	横浜N Tビル	220-0023	横浜市

"no" marker

4140794	(株)N T Tデータ		丸の内1-4-2	東銀ビルディング	100-0005	千代田区
1295340	(株)N T Tデータ		丸の内1-4-2	東銀ビルディング	100-0005	千代田区

17

8th International Conference on Information Quality, 2003

Global Customer Data Nuances ...more than meets the eye

Kanji numerals

OCRM Org Number	Org Name	Alias Name	Organization Address 1- street address (line 1)	Organization Address 1- building Name (line 2)	Organization Address 1- Postal Code	Organization Address 1- city Name
4163816	(株)N T Tデータ		中央区北一區南1-4		060-0001	札幌市
1297666	(株)N T Tデータ		中央区北一區南13-4		060-0001	札幌市

Halfwidth and Fullwidth

4184280	(株)子子子		エヌ・ティ・ティ・水田町2-11-1		060-0001	山王パークタワー100-6150
1232669	(株)エヌ・ティ・ティ	ドコモ	エヌ・ティ・ティ・水田町2-11-1		060-0001	山王パークタワー100-6150

Block data marker, Halfwidth and Fullwidth

4232882	ソニー(株)	ソニー	北区南6-7-3 5		141-8680	品川区
1269760	ソニー(株)	ソニー	北区南6-7-35		141-8680	品川区
4106261	ソニー(株)	ソニー	北区南6丁目7番35号		141-8680	品川区

18

8th International Conference on Information Quality, 2003

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.)

19

8th International Conference on Information Quality, 2003

Summary ...

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

- 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.

20

8th International Conference on Information Quality, 2003



(Gift set includes desk calculator, desk clock, and pen)

21