

Internationalization Primer

Lingoport, Inc.
3980 Broadway
Boulder, Colorado
USA 80304
+1 303 444 8020
www.Lingoport.com
info@lingoport.com

Internationalization Key Concepts

- Locale
- Translation
- Localization (L10n)
- Internationalization (i18n)
- Globalization (G11n)
- Character Sets/Encodings

Locale

- Language + territory [+ variant]
 - en_US = English (US)
 - en_GB = English (UK)
- Combines language and territorial conventions for spelling, formatting, etc.
 - en_US = "color," mm/dd/yyyy, \$1,234.56
 - en_GB = "colour," dd/mm/yyyy, £1.234,56
- A more accurate representation than language

Translation

- A rendering of the meaning of a text in another language
- Not word-for-word, but a valid representation of the intended meaning in the form in which the target language communicates such meaning
- Types of translation
 - Gisting: Provides the gist of a text in another language, useful to extract the general intent, accomplished with machine translation
 - Professional-quality: Requires a qualified human translator

Localization (L10n)

- The adaptation of a product for a particular locale
- Common abbreviation is L10n (L + 10 letters + n)
- Includes translation, also application of locale-specific behavior (formatting, parsing, etc.)
- Unless built incorporating internationalization, a product is inherently localized (specific to a given locale)
- A product must be internationalized first

Internationalization (i18n)

- The transformation of a product from locale-specific to locale-neutral
- The process of engineering a product so it can be adapted to target languages and regions efficiently and without requiring subsequent engineering changes to the core product
- Common abbreviation is i18n (i + 18 letters + n)
- A product must be internationalized before localization can occur

Globalization

- The process of transforming a locale-specific product into one that supports all target locales
- The combination of i18n and L10n

Character Sets/Encodings

- Character set
 - A set of characters used to support a given language or series of languages
- Character encoding
 - A set of code points that defines numeric values for each character within a character set (coded character set)

8-bit Character Encodings

- Latin-1 encodings
 - ISO-8859-1, Windows-1252 (Cp1252)
 - Western European languages (English, Danish, French, German, Italian, Norwegian, Portuguese, Spanish, Swedish, etc.)
- Latin-2 encodings
 - ISO-8859-2, Windows-1251
 - Central/Eastern European languages (Czech, Hungarian, Polish, Slovak, others)
- ISO-8859-3 - 16
 - Cyrillic, Arabic, Greek, Hebrew, Turkish, Baltic, etc.

Multibyte Encodings

- Single-byte characters (ASCII, Cyrillic, etc.) + double-byte characters (Asian - Kana, Hangeul, Kanji / Hanzil / Hanja)
- GB-2312 (guobiao=“national standard” in Chinese)
 - 7,445 (Simplified Chinese) characters
- Shift-JIS (“Japanese Industrial Standard”)
 - 6,355 Kanji characters

Unicode Standard

- 96,447 characters from all of the world's languages
 - Majority in the 2-byte (65,536 character) range, a.k.a. BMP (Basic Multilingual Plane)
- Primary encoding forms: UTF-8, UCS-2, UTF-16
 - UTF-8: variable length encoding (1-4 bytes)
 - Used with XML, HTML, UNIX
 - ASCII = ASCII range in UTF-8
 - UCS-2: 16-bit encoding (2-byte chars)
 - Native encoding on NT-based systems
 - UTF-16: 16-bit encoding plus surrogates (4-byte chars)
 - Supports characters beyond BMP, including less common Asian characters, musical and mathematical symbols, esoteric scripts

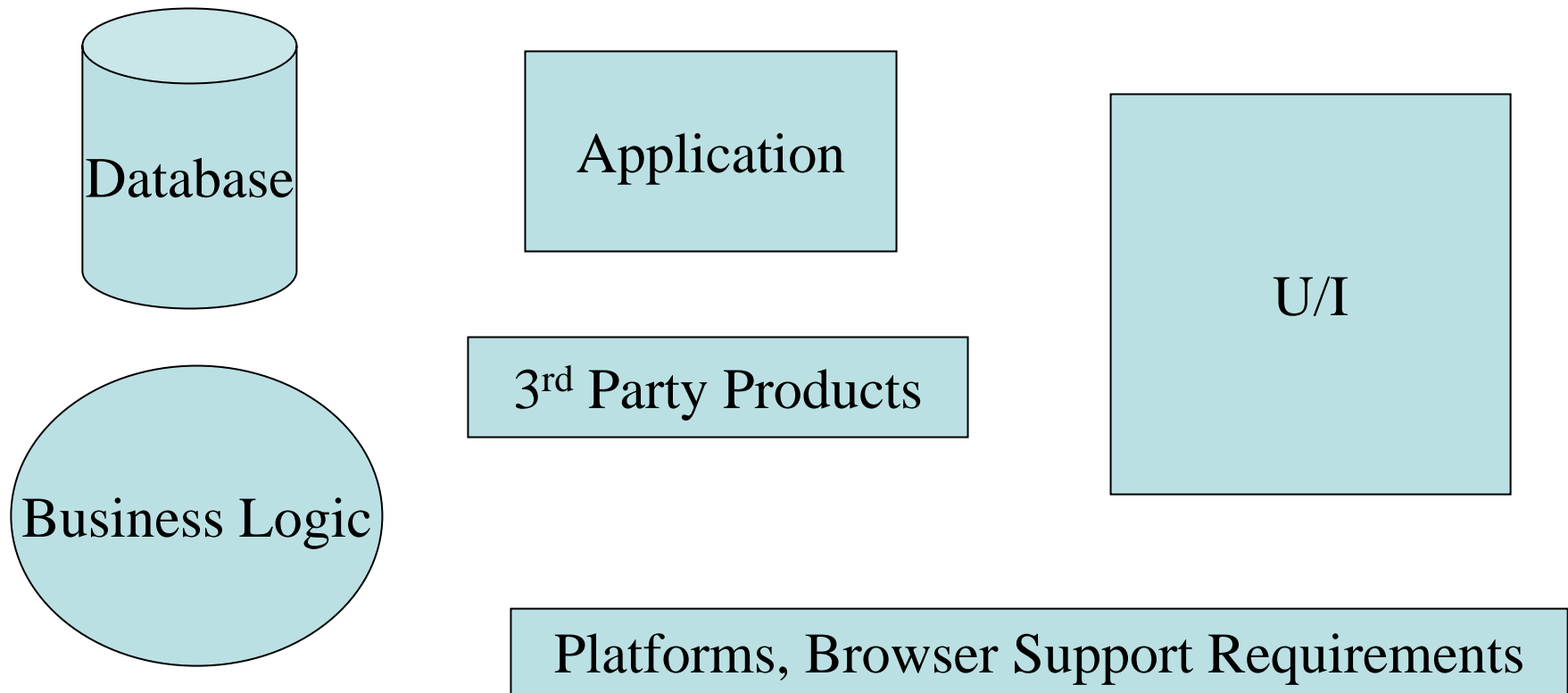
I18n Process

- Planning
- Market Requirements Analysis
- Architectural Requirements Analysis
- Code Review
- I18n Design
- I18n Implementation
- Testing
- And beyond...
 - Localization
 - Support

Market Requirements Analysis

- Target locales (languages/regions)
- Target functionality
 - Uniform across locales or locale-specific?
- Customer-driven requirements
- Overseas partner/customer feedback
- Region-specific requirements (legal, etc.)

Architectural Requirements Analysis



Architectural Considerations

- Component capabilities/functionality
 - Character encoding support
 - Locale tracking
 - Dependencies
 - Fonts
 - Service Packs
 - Libraries
 - Third Party Products
- Component interaction
 - Character encoding conversions
 - Locale notification

UI Design Considerations

- UI Layout
 - Support for string length expansion
 - Bidirectional support
 - Asian support (support for character height expansion)
- UI Locale
 - Separate monolingual vs. unified multilingual
 - Locale resolution

Code Review

- What to Identify
 - Embedded strings
 - Unsafe methods/functions
 - Image references
 - Unsafe programming constructs (ex: regular expressions)
- How to Identify
 - “Brute force”
 - Engineers search for and resolve known issues
 - Tool-assisted review
 - An I18n code analysis tool is employed to examine source code for a large range of potential and known issues
 - Issues can be identified and resolved in a more systematic fashion
 - E.g. Globalyzer (www.Globalyzer.com)

I18n Design: Key Considerations

- *Locale implementation*
 - How is locale determined, tracked and supported within the application?
- *Character encoding support*
 - What characters/encodings are required/supported?
- *Content externalization, storage and retrieval*
 - What methods are used to identify, store and retrieve translatable content?
- *UI Layout*
 - Does the layout accommodate text expansion, other locale-specific effects?
- *Locale-specific formatting*
 - Numbers, dates/times, currencies, text direction, addresses, sorting
- *Uniform vs. locale-specific content/functionality*
 - Where are the divisions between uniform content and functionality and the need for locale-specific content and functionality?

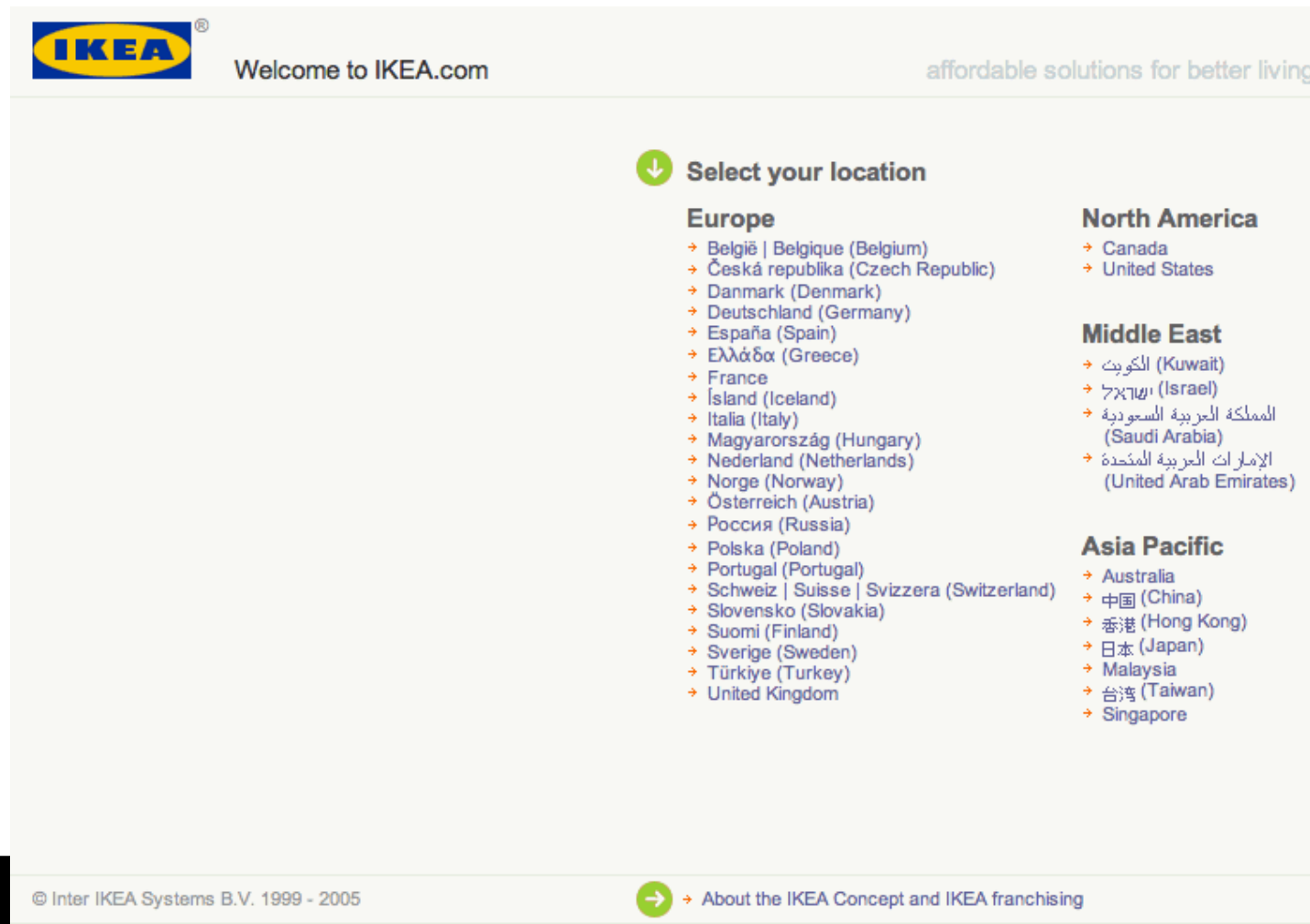
I18n Design Considerations Checklist*

- Locale implementation (determination, tracking)
- Character encodings
- Strings
 - Externalization
 - Concatenation
 - Display/Layout
- Date/time handling
- Number handling
- Currency handling
- Sorting
- Searching
- Encoding conversions
- Locale-specific functions
- Address formats
- Telephone formats
- Page layout
- Fonts and attributes
- Images, icons, colors
- Bidirectional support
- Reporting, workflow
- Database enabling
- Multi-byte enabling

Locale Implementation

- Locale determination
 - How is the current locale determined?
 - Selection process (one-time vs. global)
 - Detected
 - Preset per user/installation
- Locale tracking
 - How is the locale tracked within the application?
 - How is the locale passed between components?

One-Time Locale Selection Example:



The screenshot shows the IKEA website's locale selection interface. At the top left is the IKEA logo and the text "Welcome to IKEA.com". At the top right is the slogan "affordable solutions for better living". Below the logo is a green downward arrow icon next to the heading "Select your location". The page is divided into three columns of location options, each with a right-pointing arrow icon:


- Europe**
 - België | Belgique (Belgium)
 - Česká republika (Czech Republic)
 - Danmark (Denmark)
 - Deutschland (Germany)
 - España (Spain)
 - Ελλάδα (Greece)
 - France
 - Ísland (Iceland)
 - Italia (Italy)
 - Magyarország (Hungary)
 - Nederland (Netherlands)
 - Norge (Norway)
 - Österreich (Austria)
 - Россия (Russia)
 - Polska (Poland)
 - Portugal (Portugal)
 - Schweiz | Suisse | Svizzera (Switzerland)
 - Slovensko (Slovakia)
 - Suomi (Finland)
 - Sverige (Sweden)
 - Türkiye (Turkey)
 - United Kingdom
- North America**
 - Canada
 - United States
- Middle East**
 - الكويت (Kuwait)
 - ישראל (Israel)
 - المملكة العربية السعودية (Saudi Arabia)
 - الإمارات العربية المتحدة (United Arab Emirates)
- Asia Pacific**
 - Australia
 - 中国 (China)
 - 香港 (Hong Kong)
 - 日本 (Japan)
 - Malaysia
 - 台湾 (Taiwan)
 - Singapore

At the bottom left is the logo for "lingoport Technologies for Global eBusiness". At the bottom center is the copyright notice "© Inter IKEA Systems B.V. 1999 - 2005". At the bottom right is a green right-pointing arrow icon next to the text "About the IKEA Concept and IKEA franchising".

Global Selection Example:

18 October 2005

Selected **United States-English** >>



Federal agencies:

- » Save \$350 on select HP servers
- » See all specials

Smart Buys on handhelds for Higher education

» See all handheld specials

- » Desktops & Workstations
- » Notebooks & Tablet PCs
- » Handheld Devices
- » Monitors & Projectors
- » Entertainment
- » Music
- » Printing & Multifunction
- » Fax, Copiers & Scanners
- » Digital Photography
- » Software Products
- » Supplies & Accessories
- » Servers
- » Storage
- » Networking
- » Management Software
- » Business & IT Services
- » Solutions

» **Contact HP** » **PC Security**

» Company information » Newsroom » Register your product

» Jobs at HP » Offers/Rebates » TV ads

» **Customer Advisory – HP Lottery Ho**

[Privacy statement](#) Using this site means you accept its terms

© 2005 Hewlett-Packard Development Company, L.P.

A through B

- Africa-English
- Africa-French
- Argentina-Spanish
- Australia-English
- Austria-German
- Belarus-Russian
- Belgium-Dutch
- Belgium-French
- Bolivia-Spanish
- Brazil-Portuguese
- Bulgaria-Bulgarian

C through D

- Canada-English
- Canada-French
- Caribbean-English
- Central America-Spanish
- Chile-Spanish
- China-Simplified Chinese
- Colombia-Spanish
- Croatia-Croatian
- Cyprus-Greek
- Czech Republic-Czech
- Denmark-Danish

E through H

- Ecuador-Spanish
- Estonia-Estonian
- Finland-Finnish
- France-French
- Germany-German
- Greece-Greek
- Hong Kong-English
- Hong Kong-Traditional Chinese
- Hungary-Hungarian

I through L

Locale Detection Example:

Google™
на русском

Веб [Картинки](#) [Группы](#) [Каталог](#)

Поиск в Google | Мне повезёт!

⌂ Искать в интернете ⌂ Искать в русском интернете

[Расширенный поиск](#)
[Настройки](#)
[Языковые инструменты](#)

Google™

Web [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#)^{New!} [more »](#)

Google Search | I'm Feeling Lucky

[Advanced Search](#)
[Preferences](#)
[Language Tools](#)

Google™
العربية

[ويب](#) [صور](#) [مجموعات](#) [الدليل](#)

⌂ ضربية حظ | [بحث Google](#)

⌂ البحث في ويب ⌂ البحث في اللغة العربية

[بحث متقدم](#)
[التفضيلات](#)
[أدوات اللغة](#)

Google™
中文(简体)

[网页](#) [图片](#) [新闻](#) [论坛](#) [网页目录](#) [更多 »](#)

⌂ Google 搜索 | 手气不错

⌂ 搜索所有网页 ⌂ 搜索所有中文网页 ⌂ 搜索简体中文网页

[高级搜索](#)
[使用偏好](#)
[语言工具](#)

Character Encoding

- What character encodings are necessary to support the target locales?
- Which encodings are supported by the application components?
- Unicode is recommended in most cases, but...
 - Which encoding to choose?
 - UTF-8
 - HTML/XML, UNIX
 - UTF-16
 - Java, Windows
 - UTF-32
 - Need to support characters beyond BMP
 - It may not be supported by certain components

Strings: Externalization

- Translatable or otherwise locale-specific strings that are embedded in the source code must be externalized into locale-specific resources
- Resource files
 - Files to contain locale-specific data
 - Common file types:
 - RC files: Win32
 - Properties files: Java
 - Resx files: .NET
 - PO files: UNIX, PHP

Strings: Concatenation

- Concatenation is the piecing-together of text fragments to form a complete phrase

```
String s = "Step "+x+" of the "+proc+" process";
```

- Concatenation causes serious translation issues because of word order and grammatical differences
- Solution: use positional parameters to be replaced at runtime, e.g.:

```
String s = "Step {0} of the {1} process";
```

Grammatical Difference Example

In English

The

In German

Der Masculine Subj.
Die Feminine Subj.
Das Neutral Subj.
Die Plural
Den Masculine Dir. Obj.
Dem Masculine Ind. Obj.
Des Masculine Gen.
etc.

Strings: Display/Layout

- Fixed elements must allow for text expansion
 - Single-byte (e.g. Latin, Cyrillic, Hebrew) languages
 - expand horizontally, sometimes more than double the size of the English text
 - Double-byte character based (e.g. Japanese, Chinese, Korean) languages
 - expand vertically, since the characters are taller than Latin characters
 - Dynamic field length approach

Date/Time Handling

- Date/time parsing
 - What is 01.02.03 ?
 - February 1, 2003 in Europe
 - January 2, 2003 in US
 - February 3, 2001 in Japan
- Date/time formatting
 - Short forms have the potential to be misleading
 - Use long forms to avoid potential misinterpretation
 - ISO form: YYYY-MM-DD hh:mm:ss
 - Good for parsing, acceptable but not ideal for display

Number Handling

- Number parsing
 - How should the string “123,456” be parsed?
 - In the US, UK, Japan it is equal to the integer value 123456
 - In Europe it is equal to the float value 123456/1000
- Number formatting
 - Same question in reverse
- As numeric separators differ between locales, locale-specific functionality must be relied upon to parse and format the numbers correctly

Currency Handling

- Similar issues to number handling, but also...
- Currency formatting
 - Currency symbol/code placement
- Currency conversions?
 - For international monetary transactions

Locale-Specific Formatting Examples

Locale	Short Date	Long Date	Number	Currency
English, US	05/06/02	May 6, 2002	1,234.56	\$1,234.56
English, UK	06/05/02	06 May 2002	1.234,56	£1.234,56
French, France	06/05/02	6 mai 2002	1 234,56	1 234,56 €
Japanese	02/05/06	2002年5月6日	1,234.56	¥1,234.56

Sorting/Collation

- Q: Which list order is correct?

Ångström

Helsinki

Österreich

Zürich

Most Locales

Helsinki

Zürich

Ångström

Österreich

Skandinavian Languages

A: Depends on locale, this is just one example

Searching

- What functionality?
 - Text
 - Match all word forms: e.g. “city” vs. “cities”
 - Fuzzy matching: e.g. “security issues” vs. “security-related issues”
 - Synonym matching: e.g. “store” vs. “shop”
 - Base character matching: e.g. “société” vs. “societe”
 - Transliteration matching: e.g. “Yamamoto” vs. “山本”
 - Numbers, dates
 - Formatting must not get in the way
 - Filtering/sorting
 - By topic?
 - By date?
 - Other?

I18n Implementation

- An implementation of the international functionality requirements determined from the previous steps
- Process cycle (Globalyzer)
 - Perform code review
 - Weed out “false positives”
 - Address real identified issues
 - Test

I18n Testing

- Testing of internationalized application
 - From English user perspective, the application should function as it did prior to i18n with no new bugs introduced
- “Round tripping” of international content
 - Extended characters should be preserved in a non-corrupted state from UI to database and back

I18n Testing, cont.

- Pseudo-localization

- A “pseudo-locale” is created and implemented with “pseudo-translated” content
 - Before pseudo-translation:
UserNameLabel=Username
SomeMessage=The quick brown fox jumps over the lazy dog.
 - After pseudo-translation:
UserNameLabel=縞Ûsèèrnâàæmê史
SomeMessage=嚮Thëëë quûîck brööwn fòöõx jüüumps òvèèr thêê
lãâzÿ dõøg燭.
- Tests for:
 - String-length expansion issues
 - Extended character display/corruption issues
- Globalyzer’s PseudoJudo Utility provides a “pseudo-locale”

I18n As An Ongoing Process

- Pervasive influence on entire organization
 - Paradigm shift: not US-centric!
 - I18n Coding Standards
 - Quality Assurance
 - I18n software lifecycle tools: Globalyzer
- Avoid code forks
 - New features with international focus to launch in ALL locales
- Relationship with Localization partner

I18n Cost/Benefit

- **One-time Higher Initial Cost**
 - I18n as part of development process

- **Lower Overall Cost as End Result**
 - Future releases incur only localization costs

Questions, comments

- **What Is Your Internationalization Challenge?**
- Email or call us:
Info@lingoport.com
+1.303.444.8020

Internationalization Services: **Lingoport.com**

Internationalization Software: **Globalyzer.com**