Internationalization of Softwares

Assignment
on

Web Internationalization

Submitted for partial fulfilment of the Degree of
Masters
of
Information Technology
(Software Architecture)
Batch:2014-2016

Submitted By:
Harjot Kaur Mann
n9275398

Queensland University of Technology
Brisbane, Australia

Contents

1 Web Internationalization

1.1 First Taste: An Introduction

A terminology which is bringing awareness to cosmopolitan in the field of World Wide Web. Internationalization takes everything under the control of world. Scientifically, it enables the application to run with different vocabularies and according to their rules and customs. To arrange your web content obtainable in all other different languages and civilization for its easy accomodation for the enjoyer is known as Web Internationalization. This method of internationalization is also known as translation and localization enablement. Creating multilingual websites becomes possible with this technique, which is a powerful tool to serve the customers and buyers with better serviceability. The languages impels the text order, fonts and the accent. The crucial parts of building a website is the content and navigation but the main thing under consideration is the language which is written on the web pages and their built-in properties. There are certain specified and authorized standards which are used to represent a web site, these are Unicode. The Unicode history and the implementations for coding are provoded to the developers which are very important to take the results for building bilingual or tri-lingual websites.
It is notable here that a website possesses three folds: Content, Presentation and Behaviour.

1. **Content Layer:** The web pages contains the main content which users read when they visit the website. The text and the images are the main portion which are inhered in the content and the arrows needed by the readers for exploration of the website are also included in it. To build up the content layer, HTML is used in web development and it is also used to design the web document.

2. **Presentation Layer:** This layer presents your document to the primers. CSS handles this sheet and the approach of the document, that how it will be shown and with which media form.

3. **Behaviour Layer:** In this layer, the functioning of the webpage is achieved. For illustration, if AJAX is used, it means Javascript is enabling the functioning of the page. For backend results, Java or PHP is needed to produce the results on the demand of the user's click.

Relation to Localization

Internationalization is very much closely related to the process of Localization which is to adapt a product or website to a particular language, country or customs. This is ometimes called a locale. Internationalization and Localization are also defined by i18n and l10n respectively. This is because the Internationalization contains 18 characters between i and n and similarly in Localization, it is having 10 letters between l and n. To deine these terms more clearly, an example of is relation can be the resourceBundle class of Java. This permits for getting the locale of some information like calenders for currency. In java if resourceBundle is present there, then it means the website is internationalized.

1.2 Upbringing and Implication

According to the demonstration of Sasaki and Phillips in 2005 ans 2006 respectively, internationalization is not considered as a specific feature, it is a requirement to design the software or website in general.
Chinese believes that you can use any language to buy but you must sell in customer's language. This leads the actual demanding to plan for the multilingual and intercultural websites. In 2000, the Queens Borough Library in New York stated that the half of the habitants uses different languages at their homes to talk to each other rather than English. Due to the acquisition of this information, forces the developers to think about the internationalization of websites so that they can extend its scope to those folks who

speaks Non-English languages. But this type of translation needs much responsibility, so they need to be handle it with much care to establish their reputation in the eyes of customers that they are providing an excellent assistance and accurate assertion to their users disregarding their roots and culture. One such historical example of Internationalization is of Bosch. It inaugrated its first office of international sales in London in 1898. Its sales generation was 88 percent by 1913 over Germany.
Scott describes that In the next 10 years from 2007, there will be 70 percent of the Shopping malls will be bulit by selling to the Indians, Chinese and Russians.

How Unicode comes into play?

While designing websites, you should ensure that your website will be fitting with other cultures and languages, which leads to the existence of Unicode. This is basically a standard for ciphering industry. Implementing the persistent portrayal of content, disrespect to the script, is the main objective of this title. For examplify, there are some languages like English, arabic or Chinese, whether we write or read from left-to-right or vice versa, are indulged. Unicode uses almost ninety different scripts and billions of characters. It can be processed with differenct encodings and these are UTF-8, UTF-16 and UCS-2. The most basic character encoding style for Unicode is the UTF-8. For emails and websites, this is the default system for encoding and this will assure your website's adaptability with any language. An absolutely new term, to set the binary codes or characters. The written content of several distinct languages can be shown, handled and altered with a system called Unicode Worldwide Character Standard, which is looked after by Unicode Consortium.
Joe Becker from Xerox and Lee Collins and Mark Devis from Apple, these were the people who based unicode in 1987, by inspecting the production of Universal Character set. In the character enoding, the bytes will be generalized to characters. As the Unicode understands the scripts and languages very well and hence it manages the mulitlingual content. When Unicode is used in forms, databases and scripts, it is very effective but if you will not specify the character encodings properly, it will be difficult to read the text completely by the user.
There are many most widely used scripts which are encoded by unicode and it unifies historic and regional variations in the markup languages like XML. These variations are called glyphs. These glphys are unified into the singular characters by Unicode. Han unification is the most important example of glyphs unifiifcation, which connects the different glyphs to the variants of Chinese, Korean and Japanese to one character.

1.3 Status in Quo

Upto 2011, the third part of the world was connnected to the Internet. In Asia, the 42 percent cyberspace is residing, which is the largest population using World Wide Web while North America is 13.5 percent of the Earth's community, 24.4 percent from Europe and 20 percent from other countires including Australia. The number of people using Arabic is marked upto 1500 percent in the spent 10 years. To properly localizing the websites upto the international level, big companies have big budgets also. It means they have their own country domains for example, in India .in is used in domain to represent it as an Indian website. However, its diffcult to set their revenues upto the international endeavours for every company because it costs million dollars. So its difficult to built the websites which are internationally friendly.
So to build a world ready design, all the websites should feel locally relevant. Yunker, the co-founder of Annual Web Globalization Report Card noticed that Google is the best in building scalable products. It is obvious that Google is very strong proffessionally to lacalize its full text in all the forty languages. Without getting the full knowledge of a project, to commence a project will be a complete failure. Therefore,

it is important for the team to know the leeway and blow of the attempts. Your team shoudl divide all the areas of the application and proper analysis should be done.

1.4 Technical Come-outs

The internationalization issues of markup language and related technologies are the in the centre of interest. It is very compulsory to look at the technical issues while making a website. It is quite good to spend a quality of time specially on testing it so that you can ensure that its working good with all the devices. There are some other technical issues which you need to take in consideration.
When an application is built, every little part of it is affected rather than limiting to the presentation layer. It is concluded from the experience, that globalization of the web service has mainly two ways, and these are: Internationalization and Localization. Both these parts are foursquare to the web application layers that shown as in the diagram.
The following detail will elaborate these affairs.

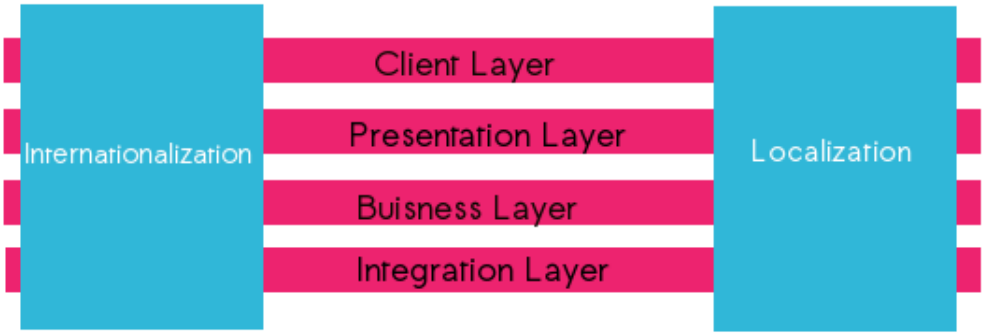


Figure 1: Internationalization and Localization orthogonal to web application layer

1.4.1 Issues related to Database Layer

A database is an important part of every website, so it is important to take care of the following points:

1. **Character Set:** The character set features include:
   - The data or content which is coming in should be sustained by the database. All the language conditions should be concerned by the character set. There are some standards which supports the English as well as some other European languages like ISO 8859, but the problem is it is not promoting Chinese, so Unicode can be used.
   - Refactorization should be there, so that there will be no need to cahnge the code and t recompile it when localized.
   - The data should be handled by the foreign incorporates.
2. **Migration of Data:** It is not diffcult to migrate the data to database because, its already encoded in the ASCII format but still to export and import data to shift the character set, there are number of tools accessible from database merchants.

3. **Width of Characters:** For every English Character, a single byte is used to encase it. So, CHAR(10) means the 10 English characters that are encoded in ISO 8859, but in Russian and Chinese language, its length may be three bytes. So for accomodating Asian and other multibyte languages, you need to give the go-ahead, so that the size of character fields can be expanded to atleast three times compared to the existing one.

4. **Buisness Induction:** Procedures and Functions incorporate Buisness Induction and the code can be refactored by doing some modifications to the sizes of the colums, so to have these changes , there are specific features provided by the database vendors.

1.4.2 Issues related to Buisness Layer

The layer which contains most of the code of the application or the site. The buisness fold matters are:
Locale Arbitration: The US English, that is en_ US locale, is beign used in the applications, which is accessible to all the users as a default case. This is actually conferencing the user's locality in the means of the world wide scenariors. The three methods in which this locale arbitration can be done are given below:

- By deducting the locale from Accept-Language HTTP header.
- For separate locales, give the different entry points.
- The locales should be accumulated according to the buyers's priority so that all desires can be based on that selections.

1.4.3 Issues related to the Presentation Layer

The way, the website looks like is the presentation of it and the layer is called the presentation layer. The user directly interacts with it and this is one which is having ultimate influence and it is having two types of contents for the user: Static and Dynamic Content.

1. **Static Content:** The static content is what that consists of pages of conditions, files, pictures etc. By managing different copies of these per language and depends on user's locale select the suitable version is the best solution. This is best handled by Content Management System and the HTTP servers such as Apache can present it outstandingly.

2. **Dynamic and Textual Content:** It refers that the website undoubtedly provides the numeric and monetary values. It permits the template groups as a unit and the contents is the resource which must be easily converted. The images and the text are treated as the dynamically generated data. By develoing the in-house tools you can extract the textual content into the resource bundles. You can use these resource bundles as the object of localization for the different languages.

3. **The Screen Design:** All languages behaves differntly while apperaing on screen, for example, if translating to the Indian Language i.e Punjabi, Hindi or any other language, if the enteties like Order Number or any Product name is not fitting to the 100 px width, then you have to get a solution for it by parametrizing the layout of the screen that is you need to do some changes with HTML div styles in the web page. In this way you can manage the design of the screen with the help of CSS. Basically you need to design a different style sheet for the languages in which you want to translate your website or application.

4. **Localizing Data Values:** The application peripheral interface generally focuses on various data lookups, For example, in case of Data fomats and Numbers the J2SE provides the localization. It includes:

   - **Date Format:** Date , Time and Calander formats vary from one to another language. To illustrate this point lets have an example of date format in international i.e English language and let us take a localized language like Punjabi(and Indian language). The standard date format is DD/MM/YYYY, MM/DD/YYYY or MMM DD, YYYY and below is shown in both Enlglish and localized in Punjabi:

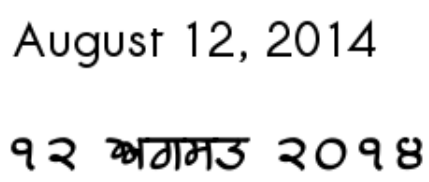


Figure 2: Date format in English and Punjabi(Gurmukhi)

Similarly, the time format in English and Punjabi is given below:

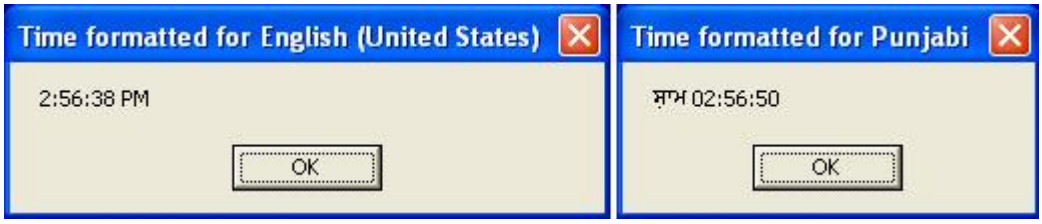


Figure 3: Time format in English and Punjabi

   - **Address and Phone number format:** An important part of Data Elements is the format of address and the phone number of an individual which differs from one country to another. For examplification, let us assume the Pin Code of Australia is XXXX and in India its XXXXXX
   - **Validations:** The variation in the format of date-time, postal codes and contact numbers, can create some problems of locale-specific validations. So, to solve these problems, there are two mathods and these are:
     (a) **Locale-specific validations:** This sype of validation is based on the Javascript and it is per-locale basis. The separate Javascript files are maintained and which are further dynamically used on the basis of localization.
     (b) **Certain Validations:** These need some server support like J2SE platform is very good in currency and date locale support. By the use of third party softwares, you can contracted for validating international services, and these implementations can be decided with the use of AJAX.
   - **Truncation of Text:** As, the sentence with same meaning can be write using different language and can vary in terms of length also, so due to the limited space in web page, a truncation scheme is used. In this method, the un-critical data or text is truncated and an option to

see the whole content will be there for user.
Let us consider a technical example of truncation in Computer Syste. In electronic form, the program can be of size 255 characters but it will alllow you to typw the maximum number of characters you want, but when the information or data will be saved, it will be truncated to 255 characters and the other characters will be considered as uncritical and it will be truncated.

- **HTTP Encoding:** In http encoding, the server will arranged a parameter called CHARSET in the HTTP header, which will define the encoding of characters in its reply. Therefore, the below code will be used in your application in JSP's.

  <code >content=text/html; charset=UTF-8 / >
  <code >

  In this, you are choosing Unicode because:

  - It allows the various languages to add in just a one page.
  - It set the CHARSET appropriately in the page and hence neglecting server-side logic.

1.4.4 Client Layer Issues:

Now, its time to take the character of Client layer into the picture. Here, the client for the application and the website is the browser. An application or a website can support mulitple browsers such as Mozilla Firefox, Internet Explorer, Chrome and many more depends on the personal liking of different kind of people and users. While developing and internationalizing your website, you team needs to take care of the following issues regarding this layer:

- **Fonts:** Font is what which is used by a browser to display the content on the screen in a specified character set. The fonts will outline the points of code to some visual representations. These font types are the part of the installation process of the operating system. To examplifiy, the Indian version of the Windows operating system will consist of Indian, i.e Hindi Language, so for viewing this language on the operating system, the font should be compatible with the machine. The Arial font of MS is an illustration of such type of such type of distributions.

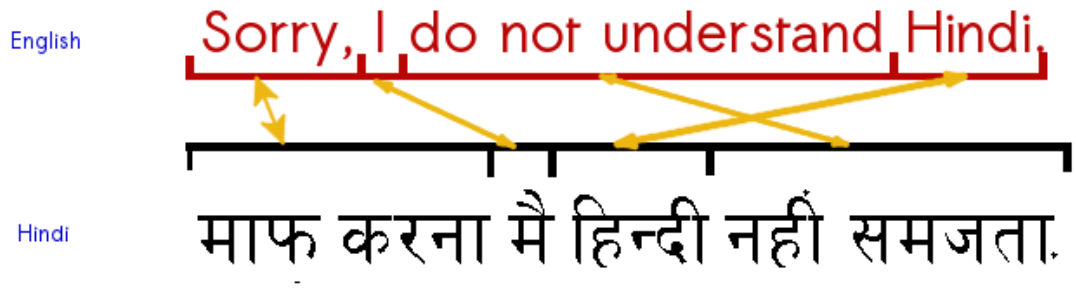
- **Javascript:** Javascript is an important client side scripting language which is also an important point to keep in mind. To perform validation and to give alert messages for the user, it is very important to use this language in an application. These should be localized as these are read before from the resource bundles Javascipt alerts are used to show it to the end-user.

- **Third party Interfaces:** In case of multilinual sites, all the third party software must be looked upon. The web services which are encoded in UTF-8 using XML handling the multilingual content. This is the XML interaction process. Flat files are those which allows the uploading and downloading of the files. This prcoss should be done carefully without losing any dat or infromation. The third party tools or API's must support multibyte character set and Unicode in which its been used. An example of this can be takes as a to generate the PDF files from your web page.

1.4.5 Linguistic and Cultural Issues:

In case of Cultural and Liguistic Issues, the Content Management Systems and Content Developers are also play an important role. This can be represented with an archetype. Lets see below:
Suppose, you are writing a sentence in an International Language and which is a combination of several words. So, it can be impervious to translate it completely to a different language style and structure. For

example, here is a sentence given below in English and it is translated into Hindi, so all the pharases can be different or reverse order. In the following diagram, a simple sentence is given which is translated into Hindi and as you can see that the order of some utterances is changed and some are at the same position. But there can be words in which no position can be same. Meanwhile, there are some cultural issues also



which have to be take care. In symbolism, which can be cultural-specific is an cultural issue. This can also be explained with the help of an illustration. In many countries, the check mark symbol means it

is correct or OK but in Japan it's meaning is that something is incorrect or not OK. They convert these check marks to circles to sybolize it for correct as localization process.
There can be many examples, but the main thing which needs to be considered in concentration is that if you localizing your website or content, it needs to be flexible, so that the process of internationalization and localization can be easily accessible and it will also be an ease for translating the text.

1.4.6 Bi-directional Text: Dreaming up a Corpus:

There are many scripts that are written from right to left, which narrows the scope of multilingual compilation to the corpora. These languages are Arabic, Hebrew, Punjabi and Hindi. Directionality, this is a property that Unicode have for each character in its rapture. This is called Bidirectional Algorithm which is for aiding the specific order of text visualization. The bidirectioanl algorithm will ensure the proper visualization, if there is only one script. For example, lets take HTML which uses Unicode as a document for character encoding and the HTML visualization of Hebrew will be WERBEH. Following is the algorithm which needs support for mixed script:

```
Source code:
 engl1 ''HEBREW2 engl3 HEBREW4'' engl5
Visualization a):
 engl1 ''2WERBEH'' engl3 ''4WERBEH'' engl5
Visualization b):
 engl1 ''4WERBEH engl3 2WERBEH'' engl5
```

The source code can be seen as:

- Two Hebrew citations with an English text or
- An Hebrew Citation with an English text, containing an English Citation.

The Unicode bidirectional algorithm will make the visualization, whithout having any other information The main cause is that this will work with embedding levels of the texts o mixed scripts i.e the Hebrew text is embedded in English language. It indictes that there can be another lavel of embedding like the Hebrew tect is contained in English text which further contains English text.
There are two methods of new embedding level. The first is that you can use the Unicode control characters as shown in the following diagram:

```
engl1 ''•U+202B•HEBREW2 engl3 HEBREW4•U+202C•'' engl5
```

In HTML, the attribute like @dir (for directionality) is used in the markup languages so that the same effect can be produced. The rtl indicates the new level of directionality from right to left or it can be ltr which is left to right like Urdu.

```
engl1 ''<span dir=''RTL''>HEBREW2 engl3 HEBREW4</span>'' engl5
```

There is an influence of directionality indicators as the plain text encompasses 27 charcters of source code while the markup languages are enclosed in 25 characters. So to avoid these affects such as queries related to word length using markup are recommended at high priority.

1.4.7 Internationalization Tag Sets

There are set of elements and attributes for intenrationalization and localization in XML and these are Internalization Tag Sets. For instance, for directionality, HTML defines and attribute. The ITS plays an important role in these sections so that their application can be nabled in the existing markup vacabularies.
A General Approach: ITS For internationalization and Localization, there are various data classes which are encompassed under the ITS Approach. Following are the categories of data that are considered under ITS specification:

- **Translate** decides whether a content or text should be translated or not.
- **Directionality** is that which tells about the information of directionality for text visualization.
- The third one is **Terminology** which is used for the foreign resources to add the reference infromation like terminology with data bases.
- Providing a means for adding important information for localization is **Localization Note.**
- W3C Ruby specification provides pronounciation by using **Ruby**.
- **RFC 4646** defines a piece of content specified by Language Information.

1.4.8 Images, Animations and Examples:

There are lot of ways to communicate with people. Its just not the text and language which leaves an impression on public mind but images, colors and objects are an important things to concentrate while localizing and internationlizing a website. You should look to the culture specific world, their way of living, body language and customs. You should alo take care of the content used in graphics or images while translating. On backgroud images, the text which needs translation can cause trouble so you should be very careful in that.

1.5 Phased Process Diagram:

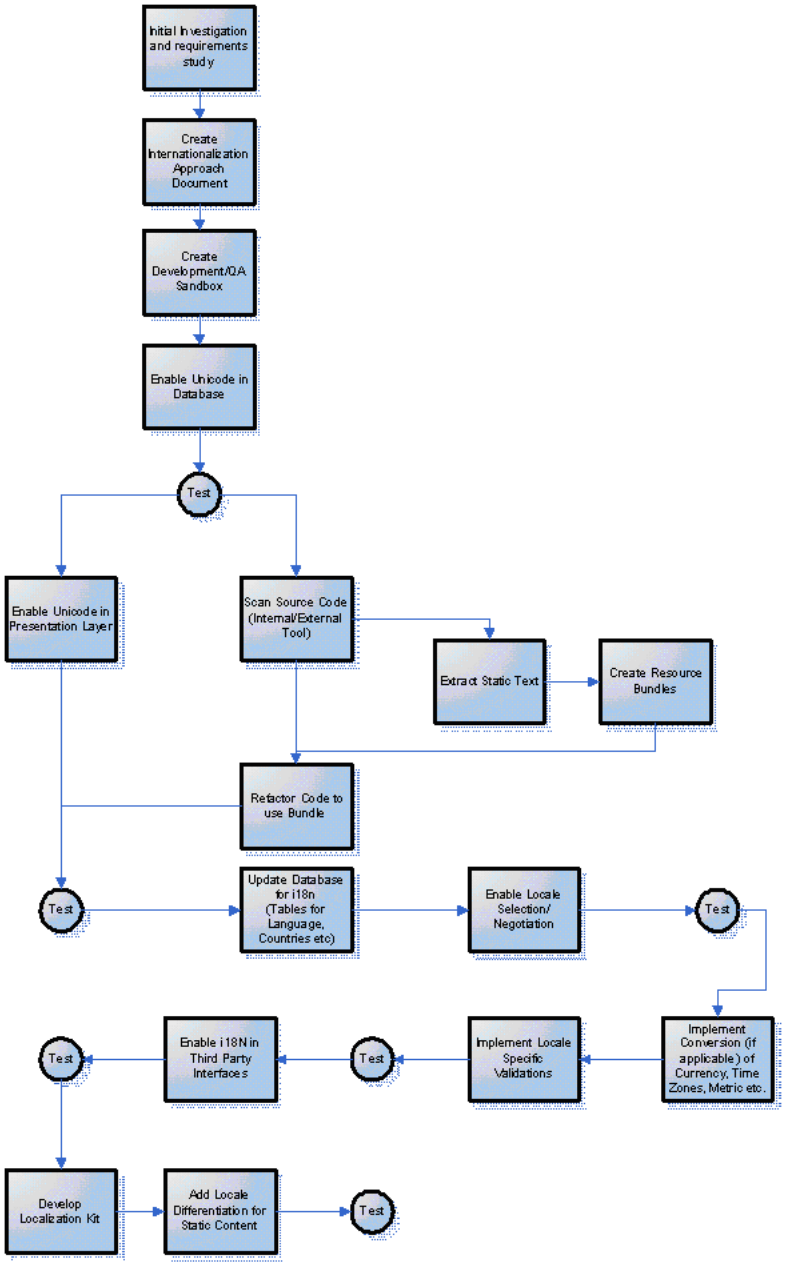


Figure 4: Phased Process Diagram

Big Bang or Phased Manner It is noticable that the process of internationalization should be in a phased manner, rather than using Big Bang approach. A phased manner is better than the Big Bang in follwing ways:

- It reduces risks, provides ease to solve the poblems fastly and efficiently.
- Regression testing can be easily done.
- Extract images and embedded content can be extracted to resource bundles.

1.6 Forthcoming Area

It is concluded from all the experiences that web internationalization is the besy method to save, represent and manipulate the data in multiple languages but its implementation is not completed and it is also not consistent in nature. The World Wide Web is the fundamental nature in all the aspects of life. It is an important point to keep in note about the popularity of Internationalizing and localizing the websites, as the languages spoken in European Union and across the world is increasing as the share of the web pages in English languages is decreasing. So in this case making multilingual websites across world wide web is extremly necessary. For this, it is important to concentrate seriously on the following issues:

- As people are not more aware of the internationalization and localization standards and practices related to the multilingual web content management. They do not understand what needs to be done.
- For enabling a multiple language websites or applications, it is important to bring the people or collaborators from different origins so that their complete involvement may help to complete the needs.
- The character aspects are the low level issues. The visualization is the musical score in the multi-layer annotations.
- The approach of ITS 1.0 will be the basis of data reuse which will project the existing markup to the new interpretations.

The data modelling, data reuse and integration of data will play an important character in the process of internationalization in the coming days but in many scenerios, the application domains can codify the needs of multilingual data. It wil also give the solution to linguistic modeling and markup languages.

References

- http://search.proquest.com.ezp01.library.qut.edu.au/docview/215831587?pq-origsite=summon
- http://blog.globalizationpartners.com/internationalization-and-accessibility.aspx
- http://www.w3.org/standards/webdesign/i18n
- http://msdn.microsoft.com/en-us/library/cc168605.aspx
- http://www.xencraft.com/training/webstandards.html#download
- http://www.bosch.com/en/com/bosch_group/history/theme_specials/internationalization/internationalization.html
- http://en.wikipedia.org/wiki/Unicode#History
- http://download.springer.com.ezp01.library.qut.edu.au/static/pdf/725/chp%253A10.1007%252F978-90-481-3331-4_4.pdf?auth66=1409574507_221c099b3baec2609b34570b498d06de&ext=.pdf
- http://www.multilingualweb.eu/about-the-project