प्रगत संगणन विकास केंद्र
CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING

L10N Guidelines & MDDS

Rajat Gupta
rajatg@cdac.in
Twitter: eRajat

Definitions

L10N	I18N	G11N

Standards

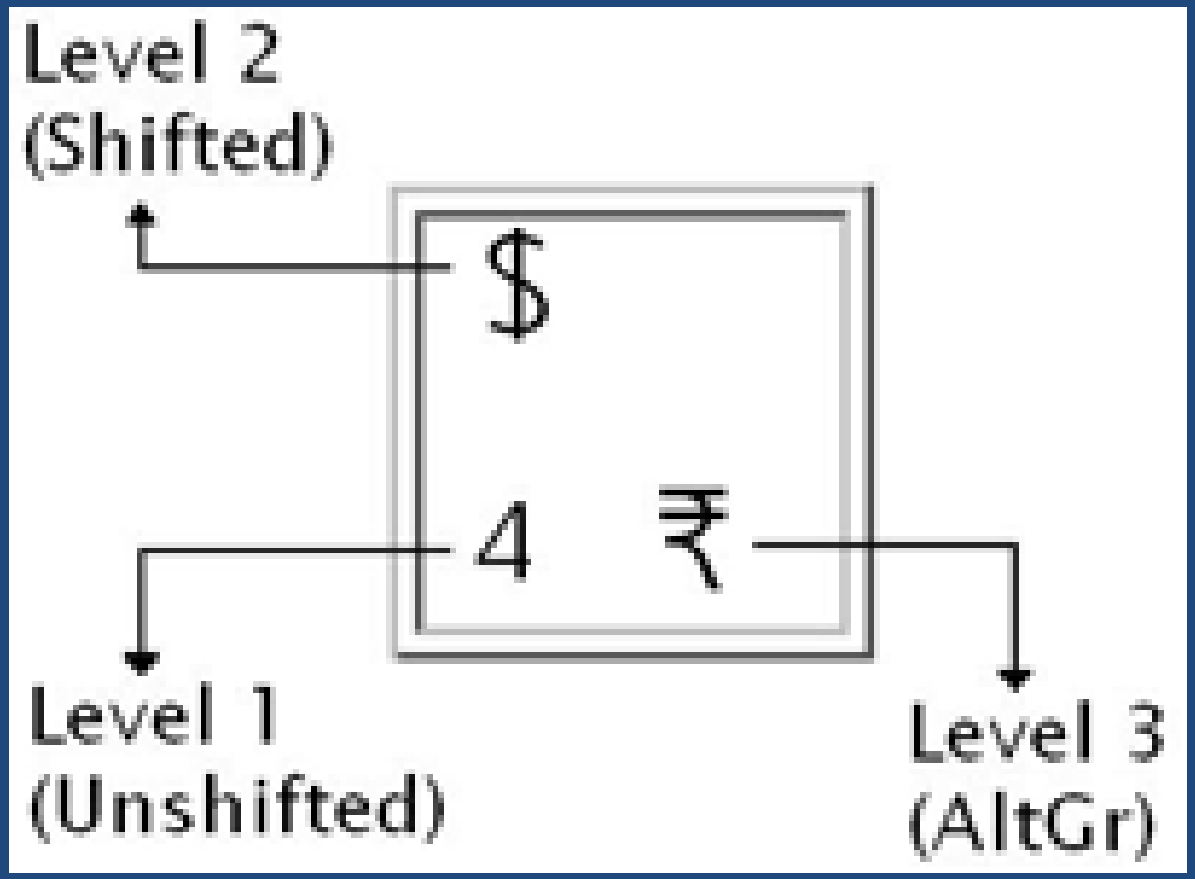
1). Text Localisation.

2). Localisation standards.

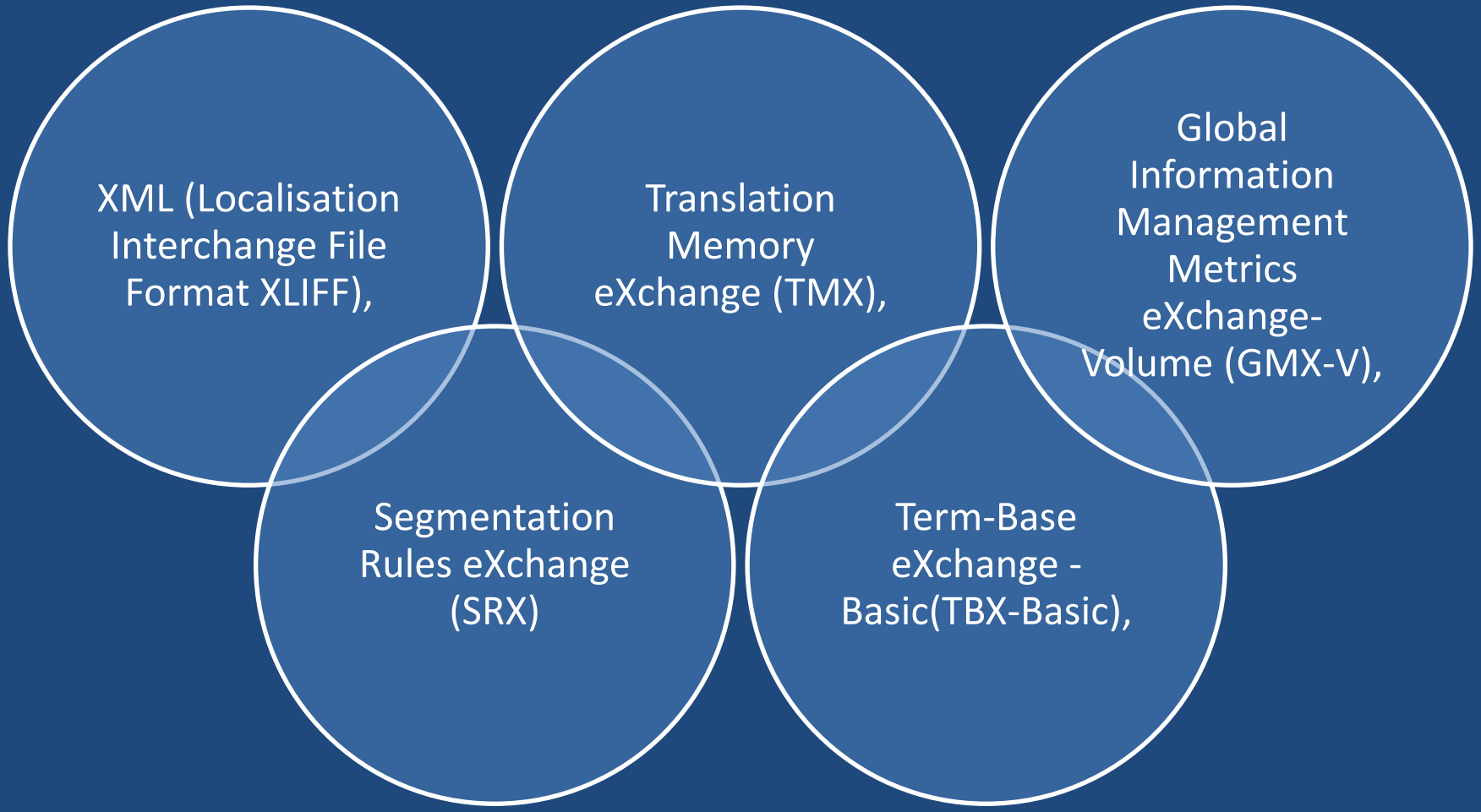
Text Localisation

- Inputting (INSCRIPT/Phonetic/Typewriter)
- Display (Font Format)
- Storage (ISCII/PASCII/UNICODE)

Rupee Symbol



Localisation Standards



Localisation Standards

XLIFF: is an XML based intermediate format which is used to store, carry and interchange localizable data.

SRX: SRX rules based on XML vocabulary was developed for breaking the text into translatable segments/ smaller fragments. SRX is defined in two parts: <languagerules>: specification about rules applicable for each language. <maprules>: specification about how rules are applied for each language.

TMX: TMX is the translation memory data exchange standard between applications. It is divided into two parts: Translation Unit <tu> and Segment of translation memory text <seg>.

Localisation Standards

TBX: TBX-Basic is a TBX compliant terminology markup language for translation and Localization processes that permit a limited set of data categories. The purpose of TBX-Basic is to enhance the ability to exchange terminology resources between users.

GMX-V: It measures the work-load for a given Localization job, not just by word and character count, but also by counting exact and fuzzy matches, punctuation symbols etc. It can also be used to count the number of pages, screenshots etc.

CLDR

- Calendars
- Numeric formats,
- Date and Time formats
- Currencies

http://cldr.unicode.org/

FUEL: Frequently Used Entries for Localization

FUEL is an open source initiative to standardize terms for open source software programs. It aims at resolving the problem of term inconsistency and lack of standardization in Computer software translation, across various platforms.

https://fedorahosted.org/fuel/

L10N Guidelines

- Refer Page 23

सी डैक
CDAC
प्रगत संगणन विकास केंद्र
CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING

CDAC Localization Guidelines MinimalSet

Point 1: Default Homepage in Marathi

All Contents should be in Marathi on Home Page.

Point 2: All Subsequent Web-Pages in Marathi.

The subsequent linked pages should be in Marathi. Many of the pages, pdf files linked are in English.

Point 3: All Menu titles of the web pages are in Marathi.

The Menu Titles must be in Marathi. We found some of the sites these are in English.

Point 4: All Web-pages developed using UTF-8 encoding.

On many WebPages the character encoding information was not found within the document of view source page. The charset attribute specifies the character encoding for the HTML document. We can declare the UTF-8 encoding in our HTML files using meta charset.

For HTML it is possible to include this information inside the head element near the top of the document:<**meta** http-equiv="Content-Type" content="text/html; charset=utf-8"> HTML5 also allows the following syntax to mean exactly the same:
<**meta** charset="utf-8"> XHTML documents have a third option: to express the character encoding via XML declaration:
<?xml version="1.0" encoding="utf-8"**?>**

Point 5: Lang attributes lang="mr" specified.

lang="mr" can be specified in the <head> tag of view source page.

Point 6: Meta tags defined in Marathi.

Meta elements are typically used to specify page description, keywords, author of the document, last modified and other metadata.

Code example of meta tags:

```
<head>
<title>Not a Meta Tag, but required anyway</title>
<meta name="description" content="मराठी">
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
</head>
```

Point 7: Are Image ALT/Captions, titles and text in Marathi.

- There should be caption in Maharashtra Emblem.
- ALT/Caption Specifies an alternate text for an image.
- The alt text should describe the image if the image contains information. The alt attribute provides alternative information for an image if a user for some reason cannot view it (because of slow connection, an error in the src attribute, or if the user uses a screen reader). A visually impaired reader using a screen reader will hear the alt text in place of the image.

- *Reference Website*: There should be proper Caption in Emblem.

  http://www.maharashtra.gov.in/en/home

Point 8: Font for the website has been provided through latest HTML5 Font SRC provisions.

With the @font-face rule, web designers do no longer have to use one of the "web-safe" fonts.

Example:

```
@font-face
{
font-family: fontName
src: url('fontFile.ttf'),//Chrome, Fire fox , Safari and Opera
url('fontFile.eot');//for IE9 browsers
}
div
{
font-family:fontName;
}
```

Point 9: All Downloads (PDF, DOC, Excel, etc.) are in Marathi language with Unicode compliance.

All the downloads should be in Marathi Language.

Point 10: All page titles in Marathi.

All Page titles should be Marathi. We found many page titles are English.

Point 11: Numbers should be in Marathi on all pages, Documents, PDFs, Images, etc.

Point 12: Contact us information given in Marathi.

Point 13: All user defined alert/error/pop-up messages are in Marathi.

Point 14: Feedback form is available should be in Marathi.

Point 15: Typing facility in Marathi is given for Interactive Website.

Point 16: Typing should be INSCRIPT layout supported.

Point 17: Onscreen Floating keyboard is available.

Point 18: Provision for increasing font size is available.

Example of provision of increasing font size of web can be seen in

Maharashtra Govt. Website. http://maharashtra.gov.in/

Point 19: In-site search support is available for Marathi language.

An in-site search is a site-specific search field. This is an internal search, which searches your website for content that matches the visitors query.

Point 20: **Website works on Hand held devices.**

We require this Information from the Website Information Manager, Since this depends on the Dept. to Dept. whether they have enabled for Handheld devices or not.

Point 21: Site map of website is in Marathi.

Sitemaps provide a way for Web sites to specify what pages within the site should be indexed and what new content has been added. Basically, it provides a communication channel between the search engine and the site. A sitemap is an XML file that contains a list of site URLs and related attributes detailing what should be indexed within a specific site. It must be UTF-8 encoded.

Reference site for Website map:

http://www.maharashtra.gov.in/web/guest/site-map

Directionality

Using the bi-Directionality algorithm, one can switch between right to left and left to right scripts.

Three languages in India viz., Urdu, Sindhi and Kashmiri are written in Right to Left direction.

The direction can be set using the following “dir” attributes.

dir = LTR | RTL

e.g.

Mohan said ") "السلام عليكمas-salaam alaykum] to me.

Directionality

Unicode Character Name	Scalar Value	Function	Equivalent Markup
LRE	U+202A	Left-to-Right Embedding	DIR attribute e.g. DIR="LTR"
RLE	U+202B	Right-to-Left Embedding	DIR attribute e.g. DIR="RTL"
PDF	U+202C	Pop Directional Format	No Equivalent </BDO> ends override
LRO	U+202D	Left-to-Right Override	BDO Element e.g. <BDO dir="LTR">
RLO	U+202E	Right-to-Left Override	BDO Element e.g. <BDO dir="RTL">

Cascading Style Sheets (CSS)

```
body { font-family: web-font, fallback-fonts; }
strong { font-family: web-font-bold; }
em { font-family: web-font-italic; }

@font-face {
    font-family: 'web-font';
    src: url('web-font.eot?') format('eot'),
         url('web-font.woff') format('woff'),
         url('web-font.ttf') format('truetype'),
         url('web-font.svg') format('svg');
    font-weight: normal;
    font-style: normal;
}
@font-face {
    font-family: 'web-font-bold';
    src: url('web-font-italic.eot?') format('eot'),
         url('web-font-italic.woff') format('woff'),
         url('web-font-italic.ttf') format('truetype'),
         url('web-font-italic.svg') format('svg');
    font-weight: bold;
    font-style: normal;
}
@font-face {
    font-family: 'web-font-italic';
    src: url('web-font-bold.eot?') format('eot'),
         url('web-font-bold.woff') format('woff'),
         url('web-font-bold.ttf') format('truetype'),
         url('web-font-bold.svg') format('svg');
    font-weight: normal;
    font-style: italic;
}
```

ITS: **Internationalization Tag Set**

ITS 2.0 is a technology to add metadata to Web content, for the benefit of localization, language technologies, and internationalization.

Metadata


metadata
use
describe
context
entities
systems
objects
definitions
ZIP
additional
model
structured
using
one
code
people
information
Data

Metadata

Metadata is key to ensuring that resources will survive and continue to be accessible into the future.

Metadata and Data Standards for e-Gov projects

MDDS: Metadata and Data Standards

Based on eGIF (e-Governance Interoperability Framework) Standard of UK

#	Item	Description
1	Name	Name / Number of the Generic or Custom Data Element
2	Description	A simple and ambiguous definition of Generic or Custom Data Element.
3	Type	Generic or Custom Generic : commonly used data element across different e-Governance applications. Custom: Used in a particular application only
4	Is Part of	
5	Parts if any	
6	Data Format	Varchar/Character/Decimal(for real/ floating number) / Integer(Whole number)/Date etc Recommended style of printing / display, if required so
7	Max Size	Maximum Size of the data element
8	Validations	Generic Validations for Generic Data and Specific Validations for Custom Data to be applied for acceptance of data.
9	Values	List of Acceptable Values
10	Default Value	For any list of values, the default value to be used unless otherwise

MDDS: Metadata and Data Standards

Name of Data Element : Gender Identification Code (G01.03)	
Description of Data Element	**Gender Identification Code of a Person**
Data Element Type (Generic / Custom)	Generic
Is part of any	
Parts if any	
Data Format	Char
Max Size	1
Validation	
Values	**M** - Male **F** - Female **T** - Transgender
Default value	
Owner	Office of RGI
Based on	-New Zealand- e Gov Standard, (http://www.e.govt.nz/Standards/e-gif/authentication/data-formats-v1.1/chapter11.html) (broken Link. Why not use ISO/IEC 5218:2004) -Census of INDIA

MDDS: Metadata and Data Standards

Name of Data Element : Marital Status (G01.04)	
Description of Data Element	Code for **Marital Status** of the Person
Data Element Type (Generic / Custom)	Generic
Is part of any	
Parts if any	
Data Format	Integer
Max Size	1
Validation	
Values	1 - Never Married 2 - Currently Married 3- Widow / Widower 4- Divorced 5- Separated
Default value	1- Never Married
Owner	Office of RGI
Based on	-Australian Govt Institute of Health & Welfare http://meteor.aihw.gov.au/content/index.phtml/itemId/291045

MDDS: iso 639-3 language codes

Recognized Official Language Code	Values	As per ISO 639-3
1	Assamese	asm
2	Bengali	ben
3	Bodo	brx
4	Dogri	doi
5	Gujarati	guj
6	Hindi	hin
7	Kannada	kan
8	Kashmiri	kas
9	Konkani	kok
10	Maithili	mai
11	Malayalam	mal
12	Manipuri	mni
13	Marathi	mar
14	Nepali	nep
15	Oriya	ori
16	Punjabi	pan
17	Sanskrit	san
18	Santali	sat
19	Sindhi	snd
20	Tamil	tam
21	Telugu	tel
22	Urdu	urd
99	Other language (English)	eng

MDDS: Metadata and Data Standards

Religion Code	Values
1	Buddhism
2	Christianity
3	Hinduism
4	Islam
5	Jainism
6	Sikhism
99	Other

Appellation Code	Values in English
1	Mr.
2	Mrs.
3	Ms.
4	Shri
11	Dr.
12	CA
13	Er.
14	Prof.

MDDS: Metadata and Data Standards

Suffix Code	Values
1	IAS
2	IPS
3	IFS
4	MBBS
5	BDS
6	MD
7	MS
8	MDS

Relationship Code	Values
1	Self
2	Spouse
3	Father
4	Mother
5	Son
6	Daughter
7	Brother
8	Sister
9	Father- In- Law
10	Mother- In- Law
11	Brother-In-Law
12	Sister-In-Law
13	Nephew
14	Niece
15	Grandson
16	Granddaughter
17	Grandfather

MDDS: Metadata and Data Standards

Ref no. of Generic data element for its Metadata	Name of Data element	Description of Data element	Data format	Maximum Size
G01.03	Gender Identification Code	M - Male F - Female T - Transgender	Char	1
G01.04	Marital Status	1 - Never married 2 - Currently married 3 - Widow / Widower 4 - Divorced 5- Separated	Integer	1
G01.05-01	Appellation Code	An **Appellation** is a title for a Person like Mr., Dr. etc. to be prefixed with the name to indicate person's gender, marital status, Professional status etc. Values as per code directory (CD01.04) ***Note: Maximum of two Appellations allowed for a person***	Integer	2
G01.06-01	Suffix Code	**Suffix** to the name of the Person to indicate person's positional status like IAS, IPS etc. Values as per code directory (CD01.05)	Integer	2
G01.07-01	Relation Type	H- Head of house hold N- Not head of household (Default value "N")	Char	1
G01.08-01	Relationship Code	Relationship of the Person, with head of the family like self, sister, brother etc.	Integer	2

Questions?

Thank You

सी डैक CDAC प्रगत संगणन विकास केंद्र
CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING

Localisation and SDLC

Tushar Kulkarni
tushark@cdac.in

Importance of L10n or I18n during SDLC

Earlier L10N had no place in SDLC

L10n becomes painful/difficult if not considered in SDLC

Sim-shipment of your products and services will be possible only if you consider L10N from the beginning itself, and make L10N integral part of SDLC.

Who is Responsible for L10N

Who is responsible for L10n

Every one at every level in the project is responsible for L10n

It's not only the job of developer but also all the stakeholders in the project

Impact L10n, I18n on SDLC

- Maintaining multiple versions of text, the programming and architecture have functional issues with L10N and I18N which makes SDLC quite different and longer
- The availability of translators and in turn translated text may affect your release plan and order of Languages to be released
- Conversely if you start thinking L10n right from the beginning and make L10n integral part of SDLC, then L10n will be straightforward. You will be in better position to localize your application and release.

Hands-On - L10N Standards

- Minimizing Impact on SDLC

Architecture decisions

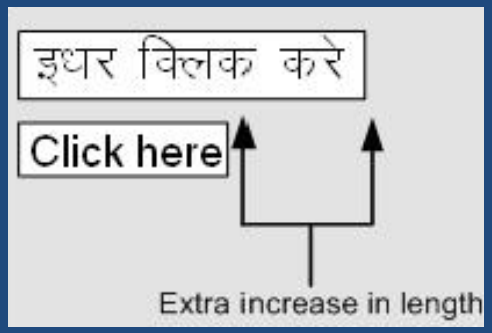
- AVOID HARD CODING :

  Avoid hard coding of strings in the project. Any display right from labels to error messages read it from a resource file.

- STRING LENGTH :

  Length of the string is also of prime importance. It's a noted fact that when we translate English language in to other language the words increase by minimum 30 to 40 %.

  For instance you can see from the below figure how the Hindi text has increased as compared to English text.



- SORT ORDER :

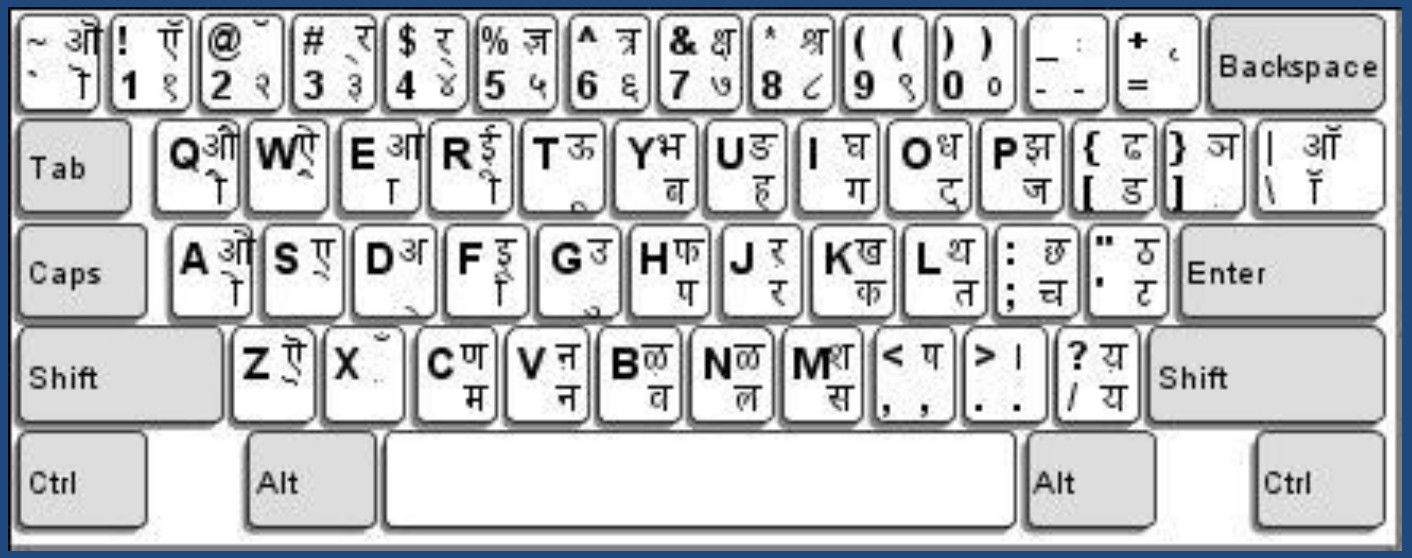
  Sort order is affected by language.

You can see from the figure below Hindi and English languages have different sorting order

KEYBOARD LAYOUT :

Keyboards layout changes according locale and region. So be careful while designing the short cut keys.

The function keys are mostly present in all key boards.



USER MESSAGES :

Validation fields , Message Box and Error Messages also need to be localized in corresponding languages.

STYLE SHEETS :

Some languages, such as Punjabi, Malayalam , Telugu are difficult to read at font sizes that are perfectly legible for languages like English .

So Using separate style sheets is a solution to this problem.

.NET Framework & Localization

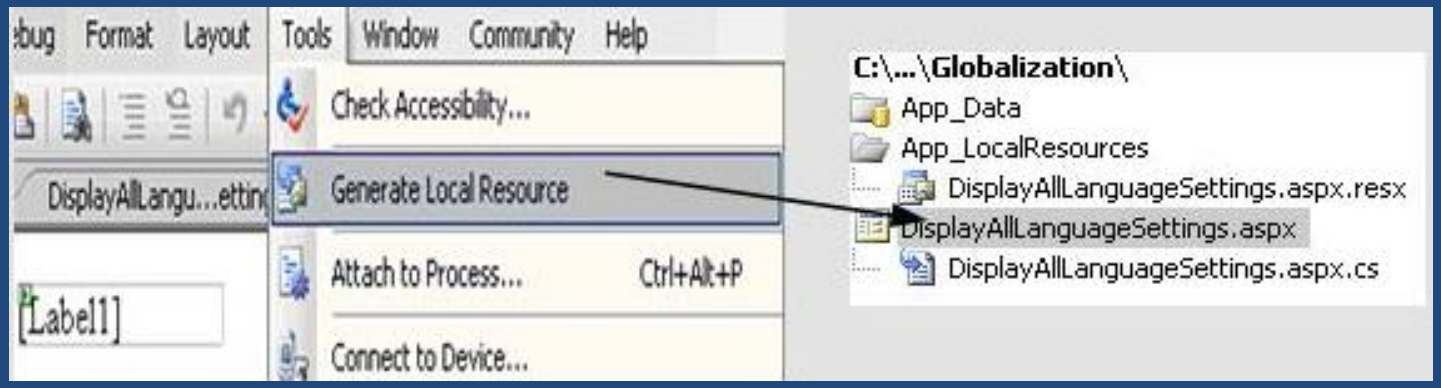
- Below is the code snippet, which shows how we can display the user languages. The first figure is the code snippet, which shows how to use "Request.UserLanguages". The second figure shows the output for the same

```
protected void Page_Load(object sender, EventArgs e)
{
    foreach(string strLang in Request.UserLanguages)
    {
        Response.Write(strLang + "<br>");
    }
}
```

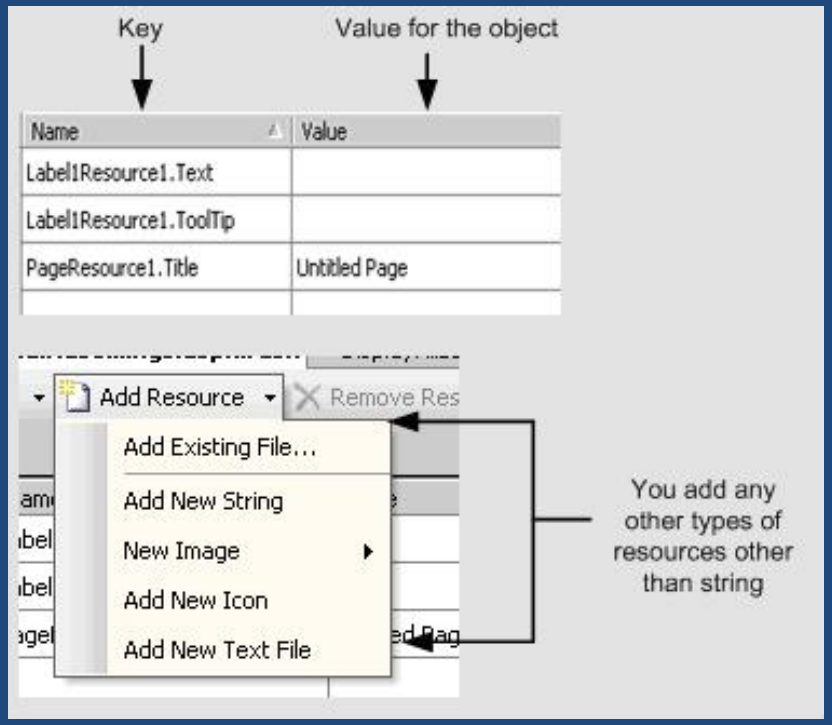
This will return a array of languages supported by the end browser.

Resource Files

- Resource files are files, which contain program resources. Many programmers think resource files for only storing strings. However, you can also store bitmaps, icons, fonts, wav files in to resource files.

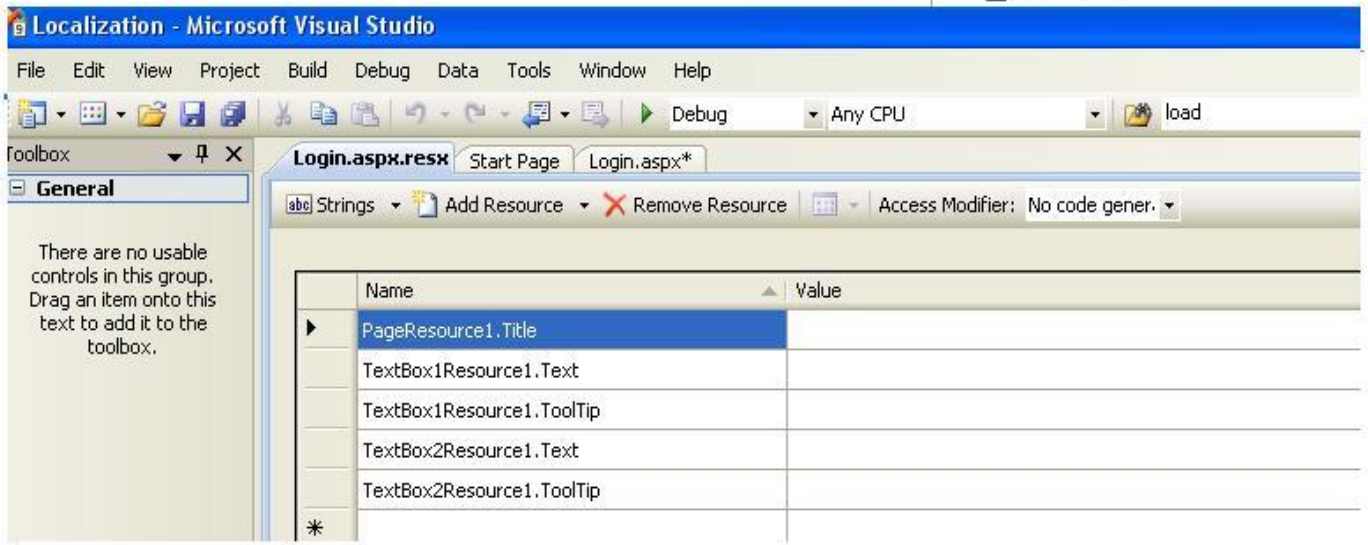


- If you see the resource file it will basically have a key and the value for the key.



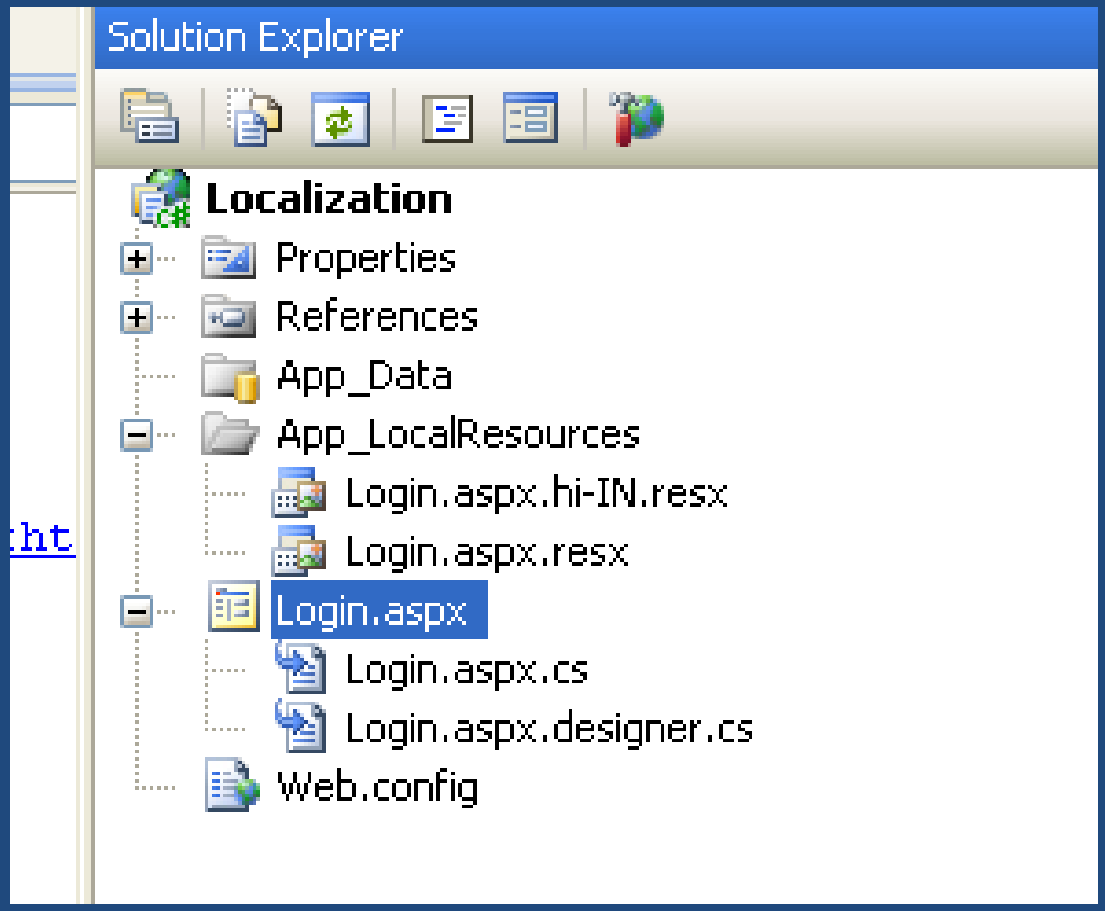
The key is basically the object name. You can see the Label1 has some value stored in the resource file

In the above figure, you can see the login page. it is a simple page with two labels and two text boxes.




```
User Name    
<asp:TextBox ID="TextBox1" runat="server" meta:resourcekey="TextBox1Resource1"></asp:TextBox>
<br />
Password      
<asp:TextBox ID="TextBox2" runat="server" meta:resourcekey="TextBox2Resource1"></asp:TextBox>
<br />
```

- Make two resource files as shown below one for Hindiand other for English. There are three values defined for "Userid", "Password" and the main title of the page. The other important thing to note is the naming convention of the files.
- You need to tag the naming convention with the language code.



- Binding Resource with selected locale

```
protected void Page_Load(object sender, EventArgs e)
{

    String selectedLanguage = "hi-IN";
    //String selectedLanguage = "en-US";

    UICulture = selectedLanguage;
    Culture = selectedLanguage;

    Thread.CurrentThread.CurrentCulture =
        CultureInfo.CreateSpecificCulture(selectedLanguage);
    Thread.CurrentThread.CurrentUICulture = new
        CultureInfo(selectedLanguage);
    base.InitializeCulture();

}
```

Languages and locale

Culture Name	Language-Country
gu-IN	Gujarati - India
hi-IN	Hindi - India
kok-IN	Konkani - India
kn-IN	Kannada - India
mr-IN	Marathi - India
pa-IN	Punjabi - India
sa-IN	Sanskrit - India
ta-IN	Tamil - India
te-IN	Telugu - India

SQL Server 2005 and International Data: Using Unicode with SQL

- Use nchar, nvarchar and ntext data types to store Indic/Unicode data
- Prefix your string literals with N (capital N – case sensitive)
- Since the Length of Regional Data is usually more than English data, while providing data type Length ,it is recommended to use nvarchar (max)

Ex :

```
SELECT * FROM TeluguDictionary WHERE (Telugu like N'%అక్క%')
INSERT INTO TeluguDictionary VALUES ('akkadi',N'అక్కడి')
```

Retrieving data using Database queries

- In Indian Languages several words have multiple correct spellings and alternate rep हिंदी हिन्दी tion forms विट्ठल विठ्ठल

Indian Language Numerals

IL numerals are not mapped to English numerals.

So a MS-SQL query :

select * from trains_table where train_no ='5312‘ ;

select * from trains_table where train_no ='५३१२‘ ;

will give different results

Technical Challenges

- Select * from Emp where City like " दिल "
- Select * from Emp where City like " दिल् "

- Give different results in MS-SQL,
- Only the second gives results for दिल्ली

Thanks

Questions?