Local Language Support

AFGHANISTAN SCHOOL ON INTERNET GOVERNANCE

AFSIG 2017

KABUL, AFGHANISTAN

SAID ZAZAI

سعید خائی

SAID@ZAZAI.CA

TWITTER: @SMZAZAI

APRIL 26, 2017

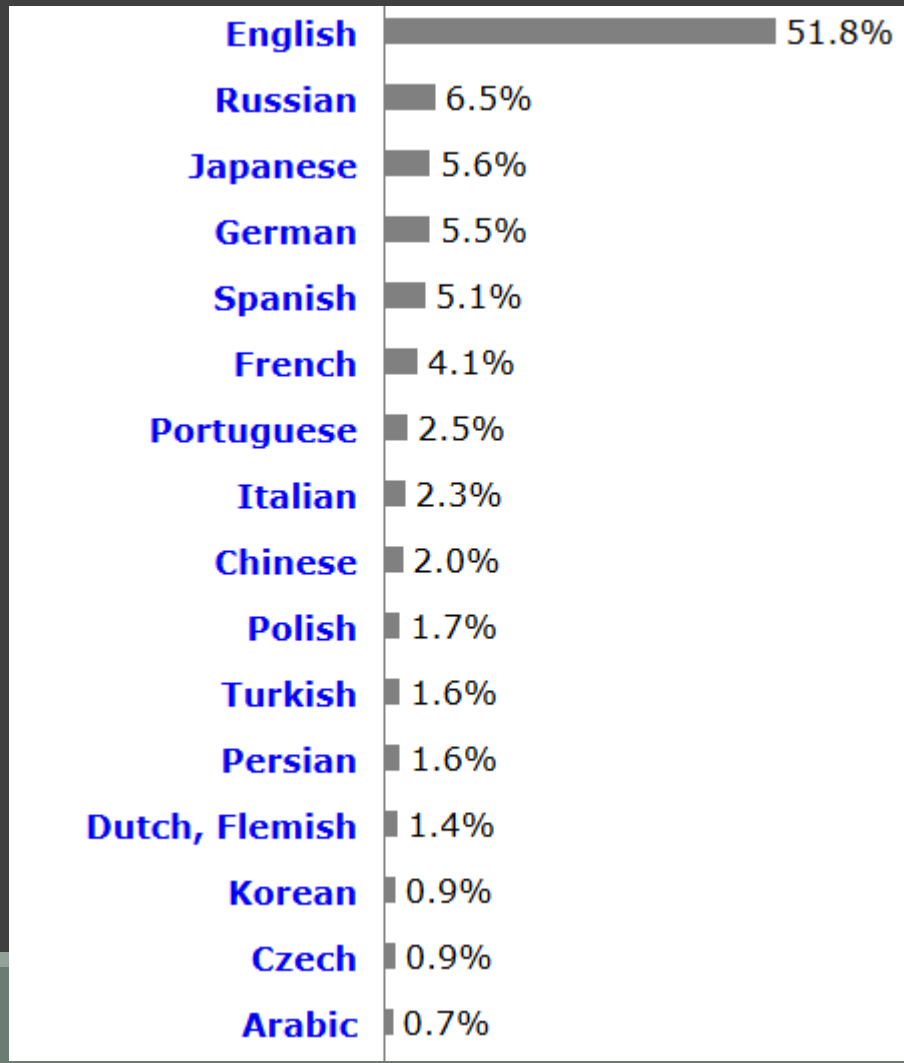
NITPAA
DIGITAL CHANGE THROUGH COLLABORATION

"Over six billion people live in over 200 countries spanning 24 time zones. These people use hundreds of currencies to conduct business in thousands of different languages and dialects. Their business practices are all over the map, ranging from simple barter to cash to electronic payment to sophisticated arbitrage."

Donald DePalma, 2002

Usage of content languages for websites

- English is used by 51.8% of all the websites.
  Source: w3techs.com
- Less than 0.1% websites use Pashto language.
  1.6% Persian (including Dari)



Communities advance when computers speak their language

IDRC reports

- A major obstacle to Internet use.
- With 3,500 local languages in the Asia-Pacific region, and fewer than 10% of people able to communicate in English, Internet use is typically restricted to urban areas.
- Highly skilled software engineers, linguists, and sociologists
  - work together to overcome the formidable technical obstacles to making local scripts compatible with computers, and to promote their use.
  - Requires a push from government, vendors, technical community, civil society and academia.

NITPAA
DIGITAL CHANGE THROUGH COLLABORATION

Local Language Support

Why need it?

- Read/Write (Community empowerment, intelligence/information society, knowledge sharing…)
- Research
- Language preservation
- Automate (Translation, text to speech, speech to text…)

Localization & Internationalization

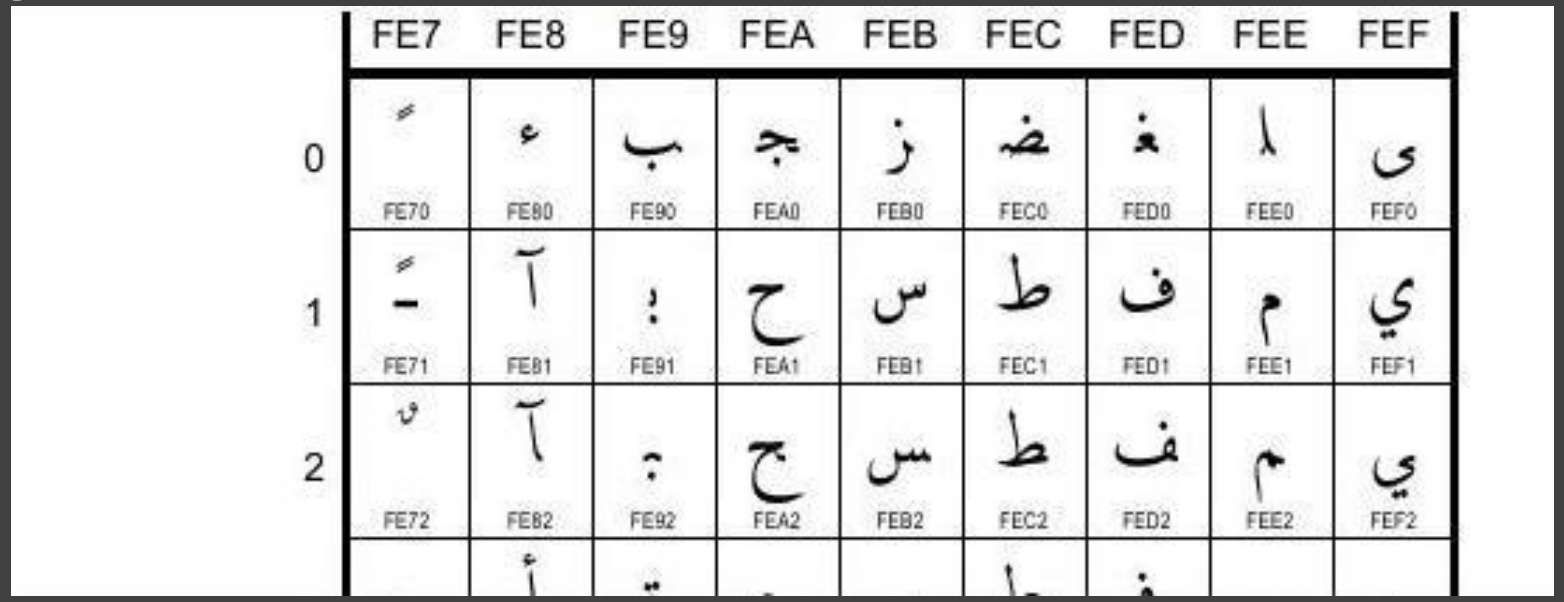
- ***Internationalization*** is the process of designing a software application so that it can *potentially* be adapted to various languages and regions without engineering changes.
- ***Localization*** is the process of adapting internationalized software for a *specific* region or language by adding locale-specific components and translating text.
- A ***locale*** is a set of parameters that defines the user's language, region and any special variant preferences that the user wants to see in their user interface.
  - Numbers, calendar, currency, keyboard layout, paper size, date-time format, phone number format, system of measurement, postal address format etc.
- It may also consider differences in culture such as:
  - Local holidays, personal names and titles, local custom and conventions, architecture etc.

The Technology

Vendor/OS/Environment	Read (Fonts)	Write (Input Method Editor)
Microsoft Windows	Windows XP	Windows XP SP3
Google Android	Android 5	Android 7
MacOS		OS Capitan
Java	Unicode	Unicode
Oracle	Unicode	Unicode

Unicode

- Unicode is a universal encoded character set that enables information from any language to be stored using a single character set.
- Unicode provides a unique code value for every character, regardless of the platform, program, or language.



	FE7	FE8	FE9	FEA	FEB	FEC	FED	FEE	FEF
0	ً FE70	ء FE80	ب FE90	ج FEA0	ز FEB0	ض FEC0	غ FED0	ا FEE0	ى FEF0
1	ـً FE71	آ FE81	ب FE91	ح FEA1	س FEB1	ط FEC1	ف FED1	م FEE1	ي FEF1
2	ٌ FE72	آ FE82	ب FE92	ح FEA2	س FEB2	ط FEC2	ف FED2	م FEE2	ي FEF2
		أ	ت			ظ	ق		

Standardization Issues

٠ ٩ ٨ ٧ ٦ ٥ ٤ ٣ ٢ ١

٠ ٩ ٨ ٧ ٦ ٥ ٤ ٣ ٢ ١

1 2 3 4 5 6 7 8 9 0

Standardization Issues

Calenders	
Hijri Solar (شمسي)	١٣٩٥
Hijri Lunar (قمري)	١٤١٠
Gregorian (ميلادي)	2017

Standardization Issues

ک ك

گ ګ

ے ي ی

Solution: Normalization

Languages of Afghanistan

Language	Alphabet Set (Written Language)	IME
Pashto	Yes	Yes
Dari	Yes	Yes
Uzbeki	Yes	No
Turkmeni	Yes	No
Pashayi	Yes (2003)	No
Sanglechi	No	No
Ishkashmi	No	No
Balochi	Yes	No
Qizilbashi	No	No
Kyrghyz	Yes	No

NITPAA
DIGITAL CHANGE THROUGH COLLABORATION

Q&A