All About Arabic *for Programmers*

▌ Overview

The Arabic script is *the* most difficult to deal with in software.

This workshop teaches the key aspects of Arabization i.e. the process of adding Arabic support to software systems. It starts with a brief overview of the Arabic culture and follows with the Arabic writing system and its contextual rules, complex ligatures, harakat (vowels), numbers and bi-directional display, etc.

The workshop then covers Arabic data representation in depth: characters sets, encodings, contextual forms, Unicode Arabic blocks and the specific encoding issues for all character classes: letters, digits, neutrals, mirrors, lam-alef, harakat, etc.

With this foundation in place, it moves on to the Unicode Bidirectional Algorithm (UBA) describing basic concepts, rules, character classes, testing tools, development issues, etc. A chapter of UBA problems and fixes is also presented.

Two more chapters focus on Input and Output. Input concerns the complexities of Arabic data entry and editing, including strange UBA side-effects and how to mitigate them. Output is about fonts, justification and other UI design issues.

A discussion of Arabic Language Processing then follows, presenting Arabic scanning, handwriting recognition, OCR, text search and other technologies.

Finally, the workshop wraps up by discussing Arabic support on the Web.

▌ Target Audience

This course is intended for software developers, web developers, testers and team leaders, or anyone involved in Arabic support that has some technical background.

▌ Benefits

After taking this workshop you will be ready to start arabizing your product. You will know the main issues and common pitfalls and you will know the solutions. You will know what requirements to consider, what changes your software or Web site or app requires, and how to implement those changes.

▌ Duration

The agenda described below is for a 1 ½ day session (with exercises, 2 days are required).

▌ Pre-requisites

This workshop presumes that attendees have already taken both days of the "All About Internationalization" workshop.

▌ Agenda

1. Arabic Culture

- History of the Arabic script, calligraphy, religion
- Arabic in today's world

2. Arabic Writing System

- Basic Arabic alphabet (abjad), Persian vs. Arabic, Romanization
- Contextual shapes, kashida (tatweel), harakat, lam-alef and other ligatures
- Numeric display, digit shapes, decimal and thousands separators, mathematics

3. Arabic Data Representation

- Logical vs. visual order, contextual forms vs. characters
- Legacy encodings (ASMO, ISO, Windows) & the Unicode Arabic blocks
- Representation and transcoding for: letters, harakat, numbers, neutrals, mirrors…

4. The Unicode Bidirectional Algorithm (UBA)

- Basic concepts: Base direction, language insertions, directional runs
- Neutral character rules and mirror characters
- Numeric classes, rules, examples
- Bidi controls: RLM, LRM, RLE, LRE, RLO, LRO, PDF.
- New Unicode 6.3 directional isolates: LRI, RLI, FSI, PDI.

5. Arabic Data Entry & Editing

- Arabic keyboard layouts and usage: selecting language and direction
- Entering text: sliding ("push-mode"), "jumping neutrals", mirrors, etc.
- The bidi cursor: selection operations with keyboard and mouse
- Quirks of bidi editing

6. Selected UBA Problems & Fixes

- Hanging neutrals, mirrors or numbers on the wrong side
- Embedded UBA problems: inside Word, PowerPoint, browsers, etc.

7. Arabic Data Display

- Fonts: harakat support, ligatures, font pairing, font attributes
- Line wrapping and justification
- Layout: directional images, graphs, bidi forms, bidi digits, 2-column bidi form…

8. Arabic Language Processing

- Arabic scanning, handwriting recognition and OCR
- Searching Arabic text: literal, stem, root, phonetic and name searches
- Overview of Arabic voice (VR & TTS), NLP, MT and translation technologies

9. Arabic on the Web

- Arabic HTML
- Arabic CSS
- Arabic JavaScript
- Arabic Email

▌ Handouts

Each attendee will receive a 250+ page booklet, with ample room for notes, complete with table of contents and glossary. The booklet is designed to serve as a practical easy-to-use reference "book" for regular use during an internationalization project.

About our Instructor – Pierre Cadieux

Pierre Cadieux is a veteran with over 30 years' experience in internationalization of software, Web sites and embedded systems. He has taught internationalization at the Université de Montréal. Pierre has been technology editor for the LISA newsletter, VP Technology at ALIS and director of technology at Bowne Global Solutions.

At ALIS, Pierre pioneered the transparent handling of Arabic and Hebrew languages and created the core bi-directional technology licensed by Microsoft.

As Director of Localization Technology at Bowne Global Solutions, he carried out research and analysis on multilingual Web sites and published the first generic model of Globalization Management Systems.

Additionally, Pierre holds a B. Sc. and M. Sc. in Computer Science.