Program Details

Monday, October 26, 2015

08:30-10:00	SESSION 1 TUTORIALS

Presenter:

Addison Phillips
Globalization Architect, Amazon

Track 1: An Introduction to Writing Systems & Unicode

This tutorial will provide you with a good understanding of the many unique characteristics of non-Latin writing systems, and illustrate the problems involved in implementing such scripts in products. It does not provide detailed coding advice, but does provide the essential background information you need to understand the fundamental issues related to Unicode deployment, across a wide range of scripts. It has proved to be an excellent orientation for newcomers to the conference, providing the background needed to assist understanding of the other talks! The tutorial goes beyond encoding issues to discuss characteristics related to input of ideographs, combining characters, context-dependent shape variation, text direction, vowel signs, ligatures, punctuation, wrapping and editing, font issues, sorting and indexing, keyboards, and more. The concepts are introduced through the use of examples from Chinese, Japanese, Korean, Arabic, Hebrew, Thai, Hindi/Tamil, Russian and Greek. While the tutorial is perfectly accessible to beginners, it has also attracted very good reviews from people at an intermediate and advanced level, due to the breadth of scripts discussed. No prior knowledge is needed.

Presenters:

Michael Ow
Staff Software Engineer, IBM

Steven R. Loomis
Software Engineer, IBM

Track 2: Putting ICU to Work

This tutorial gives attendees everything they need to know to get started with working with text in computer systems: character encoding systems, character sets, Unicode, and text processing, using the International Components for Unicode library (ICU).
ICU is a very popular internationalization software solution. However, while it vastly simplifies the internationalization of products, there is a learning curve.

The goal of this tutorial is to help new users of ICU install and use the library. Topics include: Installation (C++ libraries, Java .jar files, Java SPI for JDK integration), verification of installation, introduction and detailed usage analysis of ICU's frameworks (normalization, formatting, calendars, collation, transliteration). The tutorial will walk through code snippets and examples to illustrate the common usage models, followed by demonstration applications and discussion of core features and conventions, advanced techniques and how to obtain further information. It is helpful if participants are familiar with Java, C and C++ programming. Issues relating to ICU4C/C++ as well as ICU4J (Java) will be discussed. After the tutorial, participants should be able to install and use ICU for solving their internationalization problems.

This presentation will include newer and updated ICU features such as RelativeDateTimeFormatter, AlphabeticIndex, MeasureFormat and the DateTimePatternGenerator.

Presenters:

Tex Texin
Globalization Architect, Xencraft

Craig R. Cummings
Principal Software Engineer - Internationalization, Informatica

Michael McKenna
I18n Product Owner,

Track 3: Unicode in Action

The Unicode in Action tutorial is a 90 minute session that demonstrates programming with Unicode and related best practices.

This tutorial will build a simple application and demonstrate the code and resulting behavior as internationalization functions are added. Attendees will be able to relate these prototype examples to the requirements of their own applications and reference them to code solutions.

The program will show sorting of different strengths, regular expressions, Unicode normalization, bidirectional languages, and other features of the Unicode standard. The tutorial will highlight why each of these functions are needed so you can determine when to use them in your applications.

PayPal, Inc.

10:00-10:30 - Morning Refreshments

10:30-12:00 **SESSION 2 TUTORIALS**

Track 1: An Introduction to Writing Systems & Unicode (Cont'd.)

Presenter:

Addison Phillips
Globalization Architect, Amazon

This tutorial will provide you with a good understanding of the many unique characteristics of non-Latin writing systems, and illustrate the problems involved in implementing such scripts in products. It does not provide detailed coding advice, but does provide the essential background information you need to understand the fundamental issues related to Unicode deployment, across a wide range of scripts. It has proved to be an excellent orientation for newcomers to the conference, providing the background needed to assist understanding of the other talks! The tutorial goes beyond encoding issues to discuss characteristics related to input of ideographs, combining characters, context-dependent shape variation, text direction, vowel signs, ligatures, punctuation, wrapping and editing, font issues, sorting and indexing, keyboards, and more. The concepts are introduced through the use of examples from Chinese, Japanese, Korean, Arabic, Hebrew, Thai, Hindi/Tamil, Russian and Greek. While the tutorial is perfectly accessible to beginners, it has also attracted very good reviews from people at an intermediate and advanced level, due to the breadth of scripts discussed. No prior knowledge is needed.

Presenter:

John Emmons
Senior Software Engineer
IBM

Track 2: New! The CLDR Tutorial

The Unicode Common Locale Data Repository (CLDR) project is the largest and most extensive repository of locale data available in the industry today, providing many of the key elements necessary for proper localization of software in the various languages of the world. Since its inception more than a decade ago, the size and scope of the CLDR project has increased dramatically, as more companies and individual software developers have realized the benefits of using a common and authoritative set of data elements.

Join us for an in depth tutorial presentation, as we discuss the various types of data available in the CLDR, the data submission and vetting process, deployment strategies, lessons learned, and ways in which any Unicode member, whether individual or corporate, can participate in the CLDR project.

Presenter:

Craig R. Cummings
Principal Software Engineer - Internationalization, Informatica

Track 3: Bidi on Android and iOS

This session will an overview of the how-tos for developing native bidi Android and iOS applications. In addition to development techniques, some hints, tips, and tricks will be covered. Android Studio and Xcode development will be demonstrated live.

12:00-13:00 - LUNCH

13:00-14:30 **SESSION 3 TUTORIALS**

Track 1 - Introduction to Unicode and Beyond

Presenters:

Tex Texin
Globalization Architect, Xencraft

Craig R. Cummings
Principal Software Engineer - Internationalization, Informatica

Michael McKenna
I18n Product Owner, PayPal, Inc.

Unicode is the international encoding standard covering every major language on the planet. It is essential to know how it is designed and how to use it to be effective at text processing, handling, and debugging content. This tutorial will cover the history and creation of Unicode, its design , architecture, and examples of how it has been implemented in the real world.

The modules of the tutorial will cover:

- Why is the Unicode standard necessary? What problems does it solve?
- How computers work with text: Introduction to glyphs, character sets, and encodings.
- Unicode Standard Specification and Related Data and Content
  - Principles of Unicode's Design
  - Components of the Unicode standard
    - Encoding forms, behavior, technical reports, database
  - How to use the Unicode Standard
  - Related standards - Integration with RFCs, IETF, W3C, and others
- Unicode Implementation Details and Recommendations
  - Attributes, compatibility, non-spacing characters, directionality, normalization, graphemes, complex scripts, surrogates, collation, regular expressions and more
- Unicode and the Real World - Support for Unicode in software platforms
  - International Components for Unicode (ICU)
  - Unicode in web servers, application servers, browsers, content management systems, and operating systems
  - Programming languages JavaScript, Node.js, C/C++, Java, PHP, SQL
- How Unicode is evolving
  - Adding minority and other scripts, languages, and improving linguistic processing

---

Presenter:

Markus Scherer
Unicode Software

Track 2 - Tailoring Collation to Users and Languages

This interactive session shows how to use Unicode and CLDR collation algorithms and data for multilingual sorting and searching. Parametric collation settings - "ignore punctuation", "uppercase first" and others -

Engineer, Google, Inc.

are explained and their effects demonstrated. Then we discuss language-specific sort orders and search comparison mappings, why we need them, how to determine what to change, and how to write CLDR tailoring rules for them.

We will examine charts and data files, and experiment with online demos. On request, we can discuss implementation techniques at a high level, but no source code shall be harmed during this session.

Presenter:

Anshuman Pandey
Department of Linguistics, University of California, Berkeley

Track 3 - Introduction to Indic Scripts

'Indic' or 'Brahmi-based' scripts are currently used for the visual representation of languages spoken by more than one billion people in South, Southeast, and Central Asia. This tutorial will provide a balanced understanding of the technical and qualitative aspects of the Indic scripts of South Asia through the lens of Unicode. It will begin by discussing important themes regarding the history and diversification of Indic scripts, from Brahmi to its modern descendants such as Devanagari, Bengali, Tamil, Tibetan, Sinhala, as well as dozens of others. The tutorial will continue by describing the typological and orthographic aspects of Indic scripts by illustrating common structural features, as well as divergences. Next, the tutorial will provide an overview of major legacy and current character-encoding standards for Indic scripts, such as ISCII (Indian Script Code for Information Interchange, 1991) and, of course, Unicode. It will then discuss the Unicode model for Indic scripts and the practical aspects, advantages, and opportunities of the model. The tutorial will also provide insights into new developments in Indic scripts and the continuing trend of script invention in South Asia. Although intended for those seeking to expand their understanding of Indic scripts, experts may find the tutorial to be of interest on account of the depth and breadth of topics and scripts that will be discussed.

14:30-15:00 - Afternoon Refreshments

15:00-16:30 **SESSION 4 TUTORIALS**

Track 1 - Internationalization: An Introduction

Presenter:

Addison Phillips
Globalization Architect, Amazon

This tutorial provides an introduction to the topic of internationalization. Understand the overall concepts and approach necessary to ready a product for a global audience, including support for different languages, for writing systems of the world, and for variations in culture.

This tutorial is all new and revised for 2015.

Presenter:

Martin J. Dürst
Professor, Aoyama Gakuin University

Track 2 - Internationalization and Localization in Ruby and Ruby on Rails

Ruby is a purely object-oriented scripting language designed to make programming fun and efficient. Ruby on Rails is the groundbreaking Web application framework built with Ruby. This tutorial will help you understand the basics for internationalization and localization in Ruby and Ruby on Rails.
The tutorial will start with a discussion of how character encoding in Ruby works, and how to make the best use of it both in throw-away scripts and in long-running applications. We will show how in Ruby, all character encodings are equal, but UTF-8 is more equal than others, and should be used with preference.

Ruby on Rails also uses UTF-8 out of the box, because this is the best choice for web applications. Ruby on Rails comes with its own internationalization and localization framework. As is typical for Ruby on Rails, this framework is very simple but easily extensible. We will discuss both the basic framework as well as several helpful extensions, e.g. for handling timezones or for translating user interface texts.

The tutorial assumes that participants have some experience with programming and/or Web applications. Experience with Ruby and Ruby on Rails is a plus, but is not a condition for attending.

Presenter:

Andrew Glass
Program Manager, Microsoft

Track 3 - Building Fonts for the Universal Shaping Engine

Windows 10 includes a new shaping engine driven by Unicode data. Thus, for the first time, Windows has shaping support for all of the complex scripts in Unicode. Now that every complex script has a shaping engine, it's time for font developers to bring their talents to supporting these scripts.

This tutorial walks through the process of building an OpenType font for a complex scripts. The tutorial covers the end-to-end process of font development at a high level and pays particular focus to developing OpenType layout rules to work with the Universal Shaping Engine.

Tuesday, October 27, 2015

09:00-09:15 ***WELCOME & OPENING REMARKS***

09:15-10:00 **KEYNOTE PRESENTATION - Babel Rousers: The 900 Year Quest to Build a Better Language**

Presenter:

After a Monday full of tutorials for new attendees and those requiring a refresher, join us Tuesday morning for a keynote presentation by Arika Okrent, linguist and author of *In the Land of Invented Languages*. Arika will be illustrating the history of approaches to language invention, both ingenious and foolhardy, by

Arika Okrent
Linguist and Author of In the Land of Invented Languages

looking at particular words from these languages.

10:00-10:30 - Morning Refreshments

10:30-11:20	SESSION 1

Presenter:

Santhosh Thottingal
Senior Software Engineer, Language Engineering, Wikimedia Foundation

Track 1 - How We Built the Most Multilingual Translation System for Wikipedia

Wikipedia has a new article translation system with closely integration to its editing workflow. The system uses Machine translation engines wherever possible. It has lot of automation to assist editors to do quick translation between languages. Automatic link target adaptations, reference, image adaptations are examples. Additional translation tools like dictionaries are also provided. The system works between any 290+ languages in which Wikipedia is present. This presentation is about the interesting challenges in building such a large multilingual system and how we solved it. I will also present some interesting observations about the relationship between languages from the perspective of source-target languages.

Presenters:

John Emmons
Senior Software Engineer, IBM

Mark Davis
Chief Internationalization Architect, Google

Track 2 - Lessons Learned: How not to Represent Data in XML

In the course of CLDR development, we've ended up learning quite a bit about the plusses and minuses of using XML as a data format. We made various mistakes as we went along, and ended up investing in substantial tooling to handle the XML structure we ended up with. The choices of how to represent data in XML have a significant effect on the ease of processing, and are applicable to a wide range of applications outside of CLDR.

We'll also touch on some of the special requirements in CLDR, such as locale inheritance, and how that impacts both the XML structure and tooling.

Presenter:

Muthu Nedumaran
Founder & CEO, Murasu Systems Sdn Bhd

Track 3 - Transliterated Input Methods with Auto-correction for South Indian Languages

Input methods for Indian languages fall into two broad categories. One uses native script, which is largely based on the Inscript standard, and the other uses latin characters on the keys that produce indic letters as outputs. While Inscript is used mainly by people for whom the respective language is their first language, transliterated keyboards are becoming popular among those who use mainly English in their daily lives but do want to write in their native languages now and then without the need to learn a completely different keyboard interface. This presentation will talk about implementing transliterated input methods for South Indian languages and cover some techniques that can be employed to offer better suggestions and auto-correction features for Tamil, Telugu, Malayalam and Kannada. The presentation will also show-case a full-fledged iOS keyboard that employs the techniques discussed.

11:30-12:20	SESSION 2

Presenters:

Nick Doiron
Sr. Apps Developer, The Asia Foundation

Sora Edwards-Thro
Student, William & Mary College

Track 1 - iLoominate: Authoring eBooks in Multiple Languages

Inspired by the USAID All Children Reading challenge, a group of former One Laptop per Child teachers came together to create "iLoominate". Our project is an all-open-source, all-JavaScript app for the web and Google Chrome, which helps writers create eBooks in Haitian Creole, Nepali, and Arabic. This session covers using Polyglot.js, Wikimedia's jQuery.IME, and the PBS Kids HTML5 Storybook to create interactive, multilingual eBooks.

The full session will cover how this project has been created from the ground up, by selecting schools in Haiti, setting up a 12V solar charging system and offline server, collaborating with teachers, breaking down barriers to teaching in Haitian Creole, and producing books with the software. We will also cover challenges to supporting early readers: building phonetic word lists for Haitian Creole, and supporting 'taskil' (pronunciation guides) in Arabic script.

With presentations by Nick Doiron and Sora Edwards-Thro, core team members who deployed iLoominate in Haiti.

Presenters:

Markus Scherer
Unicode Software Engineer, Google, Inc.

Steven R. Loomis
Software Engineer, IBM

Track 2 - New in ICU

The International Components for Unicode library, or ICU, provides a full range of services for Unicode enablement, and is the globalization foundation used by many software packages and operating systems, from mobile phones like Android or iPhone all the way up to mainframes and cloud server farms. Freely available as open-source, it provides cross-platform C/C++, and Java APIs, with a thread-safe programming model.

This presentation will provide a brief overview of ICU, with emphasis on the recent updates in ICU 55 & 56, including the latest support for Unicode 8.0 and CLDR 27/28, date/time and unit formatting improvements, and other changes (see http://site.icu-project.org/download). The presentation will also touch on ICU's planned direction for future releases.

Presenter:

Alolita Sharma
Senior Engineering Manager, Twitter

Track 3 - Understanding Multilingual Tweets

Twitter can be described as a collective consciousness representing the voices of millions of people across the world. On Twitter people tweet and share information in hundreds of languages. This talk will present text processing tools that enable understanding of tweets in different languages. We will cover 'twitter-text', which is heavily used by Twitter to understand tweets in several languages. twitter-text is an open source library which defines and recognizes URLs, @mentions, #hashtags and a lot more for Unicode languages. We will also look at examples in various language families including Indic, RTL, CJK.

12:30-13:30 - LUNCH

13:30-14:20 **SESSION 3**

Presenters:

Michael McKenna
I18n Product Owner, PayPal, Inc.

Track 1 - Multilingual Transliteration - Standards, Technology, and Issues

In the expanding European Union, a majority of people speak at least two languages. In the United States, English is becoming a second language to a larger and larger segment of the population. Even so, official documents, phone books, bibliographic records, and other digital repositories may need to present native language source information (e.g., Russian or Greek) in a transliterated form to allow it to be interpreted by someone who does not speak or read that language. This paper will present a survey of issues confronting the language engineer along with solutions and available technologies. We will look at historical standards and contexts in which the standards become useless. We will then look at early work funded by the Directorate General for Research of the European Union, then on to solutions provided by Java and software libraries. We'll finish by taking a look at some commercial universal names databases and the technology behind them.

Presenter:

Rafael Xavier de Souza
Project Lead for Globalize, jQuery

Track 2 - Globalizing Modern JavaScript Apps

jQuery has changed the way that millions of people write JavaScript, believing in a world in which all web content is built on open standards and is accessible to all users.

Globalize is the jQuery library for internationalization, which is used by big corporations like Adobe and Twitter, and that is based on the Unicode Consortium standards and specifications (UTS#35) using its Common Locale Data Repository (CLDR).

Learn how to build global JavaScript applications using Globalize, having CLDR data always up-to-date, keeping code separate from i18n content, keeping code modular, and that runs in browsers and Node.js, consistently across all of them.

Presenters:

Patrick Chew
Internationalization Manager, Change.org

Pichai Saengboon
Lanna Font Developer, Center for the Promotion of Arts and Culture, Chiang Mai University

Richard Wordingham
Lanna Enthusiast

Track 3 - Tai Tham: a "Hybrid" Script that Challenges Current Encoding Models

"The Tai Tham script is one of the least understood of the mainland Southeast Asian scripts. Comprehensive research and consensus has been difficult, due to its geographic distribution amongst five regions in four countries, its status of not being a national script, and its variation in use. This paper reviews the current encoding in Unicode of the Tai Tham script by outlining its features and comparing and contrasting features of other Brahmic scripts in mainland SE Asia. In addition, issues of and for development, encoding, and rendering will be discussed, vis-à-vis the pluricentrality represented by the related spoken languages that use them. These issues will be examined in contrast from the perspective of native users' efforts in fontography; in particular, the 'LN (Lanna)' font series and its approach to addressing encoding and use will be discussed in detail."

14:30-15:20 **SESSION 4**

Presenters:

Leandro Reis
Globalization Architect, Adobe

Richard Geraghty
International Engineering Lead, Adobe

Track 1 - Case Study: Internationalization of a World-Class Enterprise CMS

At the core of Adobe's $1B Marketing Cloud business is Adobe Experience Manager (AEM), a leading enterprise content management solution used by many of the world's top multinational corporations (e.g. Chevron, General Motors, Philips) to create multilingual web sites. Learn how two veterans from Adobe's Globalization team, Leandro Reis and Richard Geraghty, have worked together in the last 2 years to ensure AEM's internationalization by directly resolving over 600 internationalization bugs, and by working with core architects and developers to design and adapt the product so as to prevent such large number of issues from happening in the first place.

As a very large, complex and ever-growing product offering a multitude of features such as digital asset management, web analytics, social community management, e-commerce, mobile publishing, and marketing campaign management, AEM requires constant internationalization work.

Based on their internationalization work on AEM's extensive feature set, they will share implementation examples of unicode enablement, date/time/calendar formatting, multilingual searching, language-specific sorting, expandable UI layout, proper language tag usage, bi-directional text support, UI mirroring, and security-aware string externalization, using a variety of programming languages, libraries and formats

(Java/JSP/Servlets, JavaScript, JQuery, Moment.js, Handlebars.js, HTML, CSS, XML, JSON) in a multi-platform (MacOS, Windows, Unix), and multi-framework (Apache Sling, Apache Jackrabbit/Oak, Apache Felix, Adobe Granite, Adobe CoralUI) environment.

Presenters:

Craig R. Cummings
Principal Software Engineer - Internationalization, Informatica

Tex Texin
Globalization Architect, Xencraft

Track 2 - Comparing Java Script Libraries

Which JavaScript library is best for international deployment? This follow up to last conference's session presents the results of further investigation of the features of several JavaScript libraries and their suitability for international markets. We will show how the libraries were tested and compare the results for: Dojo, jQuery, Closure, iLib, and FormatJS, in addition to ECMA-402. The results often surprise and will be useful to anyone designing new international or multilingual JavaScript applications or supporting existing ones.

Presenter:

Norbert Lindenberg
Founder, Lindenberg Software LLC

Track 3 - Bringing Balinese to iOS

While today's main operating systems include good support for the most popular writing systems, support for historical and minority writing systems is often poor, unavailable, or left for third parties to provide. In the case of iOS, it recently became possible for third party apps to provide the core of writing system support, fonts and keyboards, for systemwide use. Using Balinese as an example, this presentation discusses how to implement fonts and keyboards for a complex writing system for iOS.

15:20-15:50 - Afternoon Refreshments

15:50-16:40 SESSION 5

Presenter:

Tex Texin
Globalization Architect, Xencraft

Track 1 - Agile Internationalization User Stories

User stories are the way that Agile Methodology describes the functionality of the software being developed. Each story describes an action or need of a user and in so doing defines the functions the software must provide and the requirements it must satisfy.

This session will describe the mapping of an internationalization checklist into a suite of user stories that are used in internationalizing a software project.

Presenter:

Steven R. Loomis
Software Engineer, IBM

Track 2 - Globalization node.js

Node.js has become a popular platform, using JavaScript on the server instead of its traditional role in web browsers. This presentation will discuss challenges and lessons learned in enabling the Intl (EcmaScript-402) module by default in the recent node v0.12 release, what's next for JavaScript and node.js globalization, and discuss techniques and best practices for Unicode and international support in node.js applications.

Presenter:

Hohyon Ryu
Senior Software Engineer, Twitter Inc.

Track 3 - Korean Text Processing with Twitter-Korean-Text

Spoken by about 80 million speakers worldwide, Korean is one of the most complex languages to process digitally. Korean uses its own unique writing system - Hangul. Each Korean character is composed of consonants and vowels. In some cases, a character can be composed of 2 or more morphemes. Korean language requires complicated segmentation as spacing rules are not observed strictly, postpositions are glued to nouns, and verbs can have an unlimited number of conjugations. Short and casual text on Twitter also introduce unique challenges as users casually alter words to shorten their length or to introduce new slang expressions. This session will cover a brief introduction to Korean language and Hangul, challenges in Korean text processing, and an open-source Korean text processor named twitter-korean-text.

Twitter-korean-text is an open-source project designed to address various Korean text processing issues featuring tokenization, stemming, normalization, and phrase extraction. The source and documentation is available at https://github.com/twitter/twitter-korean-text. It is designed to handle various types of Korean text including formal, casual, short, and long text.

16:50-17:40 SESSION 6

Moderators:

Alolita Sharma

Senior Engineering Manager, Twitter

Martin J. Dürst
Professor, Aoyama Gakuin University

Track 1 - IUC 39 Lightning Talks

This session will be a series of lightning talks of 5-10 minutes each. The talks should be related to internationalization, localization and any other related areas listed in the CFP topics. Alolita Sharma and Martin Durst will be moderators for the session. Questions on any lightning talk will be at the end of the 60 minute session.

See en.wikipedia.org/wiki/Lightning_talk for details.

Presenters:

Shervin Afshar
Netflix, Inc.
and
Behnam Esfahbod
Facebook

Mark Davis
Google, Inc.

Jim DeLaHunt
Jim DeLaHunt & Assoc.

Chris Dillon
University College London

Ben Hamilton
Facebook

James Koval
Twitter

Muthu Nedumaran
Murasu Systems

Anshuman Pandey
UC Berkeley

Presenters:

Michael McKenna
I18n Product Owner, PayPal, Inc.

Reza Payami
Internationalization Engineer, PayPal, Inc.

Track 2 - Griffin - PayPal Node.js API for i18n

In late 2013, PayPal made a corporate decision to migrate to Node.js. However, Node.js had very little or broken internationalization support, mostly inherited from client-side JavaScript which relied on the underlying operating system for i18n support. Initially, Node.js did not have the metadata or features needed to come even close to what PayPal needed to support markets in over 200 countries, plus appropriate English and local language formatting for every region.

This talk will discuss the effort the PayPal Internationalization Technology Team went through to investigate all known Open Source options, decide on an appropriate infrastructure to build upon, then customization of CLDR and other data to fit in to that infrastructure to finally enable Node.js products to support validation, normalization, display, and html and semantic markup of dates, times, numbers, currencies, phone, postal address, and personal names in 200+ countries and 27 languages. Included will be methods used to reduce the memory footprint, the decisions made to use regional territory containment and choice of English locales to use for fallback in the many countries where CLDR does not have an appropriate English locale defined.

As a finale, we will demonstrate the Griffin Reference App created to allow content owners to view formats according to locale, and encourage just-in-time learning by providing a pre-populated Node.js playground, similar to W3School's "Try it Yourself" JavaScript playground app.

Presenters:

Sharon Correll
Software Engineer, SIL International

Martin Hosken
Script Research and Engineering, SIL International

Track 3 - Using Graphite to Address Challenges in Nastaliq-style Arabic Script

Nastaliq-style Arabic is one of the most complex forms of writing in the world, and standard font technologies, including OpenType and Graphite, are not quite up to the challenge of handling its sloping, calligraphic form. For this reason, SIL's smart-font technology, Graphite, is being extended with some special capabilities to address the particular challenges of this beautiful but complicated form of writing.

There are two main challenges that arise in developing a Nastaliq font. One is the sheer volume of glyphs that are required, due to the complexity of the calligraphic shapes. Unlike other forms of Arabic, which require initial, medial, final, and isolate forms for dual-connecting characters, most Nastaliq letters potentially require a separate form to precede every other letter of the alphabet, in both initial and medial contexts. This means that the number of glyphs required in the font is at least O(n^2) on the number of base characters.

The sloping nature of Nastaliq creates a second, even greater challenge: glyph collisions. A straightforward, naive layout of base glyphs, nuqtas, and diacritics will inevitably result in a rendering where the glyphs collide, forming ugly and even unreadable text. Fixing these collisions is exacerbated by the large number of glyphs and the complex positioning created by the sloping baseline.

Workarounds to current font technologies have been used to create Urdu-specific fonts, but these approaches do not scale well when multiple languages and a variety of diacritics are needed. For this reason, SIL International is developing a font called "Awami Nastaliq," specifically intended to support lesser-known languages of west Asia, and using an extended version of Graphite.

To solve the problem of collisions, we are enhancing Graphite with an automatic collision-fixing capability. The Graphite engine makes use of a simplified form of the rendered glyphs to detect collisions, shift nuqtas and diacritics, and add kerning to create nicely laid-out text. Besides fixing collisions, the kerning mechanism can also create diagonal overlaps in the sloping text, as Nastaliq is traditionally written. The behavior of the algorithm can be fine-tuned using parameters defined in GDL, the programming language used for creating a Graphite font.

We expect that building collision fixing directly into our rendering engine will provide a level of power not found in any other unguided font technology.

18:00-19:00 - CONFERENCE RECEPTION

Wednesday, October 28, 2015

09:00-09:50 | **SESSION 7**

Track 1 - How the Unicode Consortium Works (And How You Can Get Involved)

Presenters:

Deborah Anderson
Technical Director, Unicode Consortium

Lisa Moore
Vice President, Chief Financial Officer, & UTC Chair, Unicode Consortium

Rick McGowan
Technical Vice President & IUC Conference Chair, Unicode Consortium

The workings of the Unicode Consortium can appear mystifying to outsiders. This panel will look behind the Unicode curtain to reveal how the organization is run, how decisions are made, how to provide input, and how to get involved.

Topics will include:

- the location, staff, and officers of the Consortium
- committees that make up the Unicode Consortium
- projects which are a part of the Unicode Consortium
- levels of membership and voting rights of each level
- how characters get approved in the Unicode Standard (emoji and non-emoji)
- how to provide input and feedback
- academic participation in Unicode activities

The panel will be made up of the Unicode Consortium CFO and UTC Chair Lisa Moore, Unicode Technical Vice President Rick McGowan, and Unicode Technical Director Debbie Anderson.

Track 2 - What Toys are in Your Internationalization Toolbox?

Presenters:

Michael Kuperstein
Localization Engineer, Intel

Loïc Dufresne de Virel
Localization Strategist, Intel

Join us for a fun session where we'll outline a systematic process for internationalizing software. We'll illustrate the process with stories of cringe-worthy disasters, near-misses, and successful product launches. From the old stand-by reviews and assessments to the flashy new wonders of automation, everyone has a favorite toy, and each has its place in the overall process. You'll gain some practical knowledge of the types of tools and processes that are available in the industry to help you develop World-Ready applications with minimum trouble and maximum fun.

Track 3 - Universal Shaping

Presenters:

Andrew Glass
Program Manager, Microsoft

Behdad Esfahbod
Internationalization Software Engineer, Google Inc.

John Hudson
Co-Founder, Tiro Typeworks Ltd.

Roozbeh Pournader
Internationalization Engineer, Google Inc.

At IUC 38, Microsoft unveiled the Universal Shaping Engine (USE). USE is a new OpenType script engine available in Windows 10 that is driven by Unicode data. It currently supports 45 complex scripts encoded Unicode and is designed to be easily updatable as new complex scripts are added to Unicode in the years to come. In this joint presentation, we discuss development of the USE engine from multiple technical angles:

- Challenges in deriving OpenType shaping from Unicode data
- Directing Unicode data development by rendering use
- Challenges and lessons learned, in implementing USE in HarfBuzz
- Beyond shaping: universal typographic sophistication

This is a 2 part session.

10:00-10:50 | **SESSION 8**

Track 1 - Emoji Q&A Panel

Presenter:

Mark Davis
Chief Internationalization

Who chooses these mysterious characters? Find out how, beneath a seemingly-ordinary street in Zürich, deep in the vaults of Gringotts, a shadowy cabal meets to decide the future of emoji.

Architect, Google

Presenters:

Addison Phillips
Globalization Architect, Amazon

Track 2 - Adventures in Android Message Formatting

One of the most basic internationalization tasks is the formatting of numbers, dates, strings, and other values in a culturally acceptable manner. Most platforms provide APIs to simplify this for developers (and maybe also for translators).

Problems can arise, though, as user interface designers create richer experiences on smaller screens. This session reviews the Java, ICU4J, and Android APIs designed to serve these requirements--through the lens of real life developer crises.

Presenters:

Andrew Glass
Program Manager, Microsoft

Behdad Esfahbod
Internationalization Software Engineer, Google Inc.

John Hudson
Co-Founder, Tiro Typeworks Ltd.

Roozbeh Pournader
Internationalization Engineer, Google Inc.

Track 3 - Universal Shaping (Cont.)

At IUC 38, Microsoft unveiled the Universal Shaping Engine (USE). USE is a new OpenType script engine available in Windows 10 that is driven by Unicode data. It currently supports 45 complex scripts encoded Unicode and is designed to be easily updatable as new complex scripts are added to Unicode in the years to come. In this joint presentation, we discuss development of the USE engine from multiple technical angles:

- Challenges in deriving OpenType shaping from Unicode data
- Directing Unicode data development by rendering use
- Challenges and lessons learned, in implementing USE in HarfBuzz
- Beyond shaping: universal typographic sophistication

This is a 2 part session.

10:50-11:10 - Morning Refreshments

11:10-12:00 | **SESSION 9**

Moderator:

Steven R. Loomis
Software Engineer, IBM

Panelists:

Shawn Steele
Senior Software Engineer, Microsoft

Mark Davis
Chief Internationalization Architect, Google Inc.

Shervin Afshar
Localization Engineer Netflix, Inc.

Shawn (Xiang) Xu
Internationalization Engineer, Netflix

Nova Patch
Lead Engineer, International Search, Shutterstock

John Emmons
Senior Software Engineer IBM

Track 1 - CLDR Users Panel

After the character properties in Unicode itself, access to language and region-specific locale data is the next most popular data needed by globalized applications. This is why the Common Locale Data Repository (CLDR) was long ago spun out from one such application library's source code. This presentation will start with an introduction to CLDR and what's new in versions 27/28, and then continue with a panel discussion focused on the experience of direct consumers of CLDR data. Topics discussed will include how to use the data and what issues have been encountered using LDML and JSON format CLDR data. This discussion will also include ample time for general Q&A about CLDR.

Presenter:

Joel Sahleen
Sr. Globalization Engineer, Adobe

Track 2 - Internationalizing and Localizing Single-Page Web Applications

In this session, we will look at the challenges, resources and strategies associated with internationalizing and localizing SPAs. The goal is not to come up with a definitive solution to all possible problems, but rather to make the audience aware of the different options available so they can make an informed choice between them. As part of our discussion, will examine how internationalization and localization are handled

in four popular SPA frameworks: Angular.js, Backbone.js, Ember.js and Meteor.js. We will also investigate some popular JavaScript i18n libraries and show how the frameworks mentioned above can be used in conjunction with them. The session will conclude by analyzing a real-world example of SPA internationalization and localization and the lessons learned from its construction.

Presenter:

Raph Levien
Software Engineer, Google, Inc.

Track 3 - **A Tour of Android Typography**

Material Design is a flagship feature of the Android Lollipop release, and typography is a cornerstone of that effort. This talk will survey recent developments in typography in support of Material Design and for Android in general, and will present technical and artistic aspects, as well as guidance for developers to achieve high quality typography on mobile devices.

A major theme is support for a large fraction of the world's languages. The Roboto 2 font features a huge increase in coverage for Latin, Greek, and Cyrillic, including many characters defined in Unicode 7. Complementing Roboto is the Noto family, supporting approximately 30 scripts in the latest Android release. Of particular note is Noto Sans CJK, again with vastly expanded coverage, including both Simplified and Traditional Chinese, Japanese, and Korean. Noto Sans CJK is also available in a wide range of weights.

Android's text stack is similarly becoming more sophisticated, giving the system and app developers access to the expanded weights and styles, typographic refinements such as letter spacing and OpenType features (including old style figures), and improved line breaking, especially for non-Latin scripts.

Android is open source, and both the fonts and the code implementing Android's typography are available for adapting into other projects.

12:00-13:00 - LUNCH

13:00-13:50 | **SESSION 10**

Track 1 - Enhanced Umm AlQura Calendar Support

Presenters:

Mohamed Mohie
Globalization Manager (PMP®), Manager of Arabic GCoC, IBM Worldwide

Ramy Said
Advisory IT Specialist, Globalization Center of Competency - Arabic Focal Arabic Competence and Globalization Center (ACGC) Cairo Technology Development Center (CTDC) - IBM Egypt

Umm AlQura calendar is used by many users from Saudi Arabia, Bahrin and Qatar in iddle East region.The previously supported Umm AlQura calendars were not accurate and shown wrong dates as reported by many developers. By the cooperation with the Official Calendar of Kingdom Saudi Arabia site http://www.ummulqura.org.sa/ and the supervisor of Management research and development projects Office in KASCT we had access to the accurate Umm AlQura year/month data table. With the aid of this data we were able to create an accurate implementation of Umm AlQura calendar and an optimized Umm AlQura/Gregorian date converter. The new approach was supported by IBM ACGC team in ICU and Dojo frameworks. The presentation will discuss how the new calendar approach was implemented and will show a live demo.

Presenter:

Jim DeLaHunt
Principal, Jim DeLaHunt & Associates

Track 2 - Building Localization Capacity Through Non-specialist Developers

Compared to 25 years ago, by any reasonable standard, we internationalization engineers have won. Unicode is the dominant character encoding, platforms have strong i18n support, global product sales are common, localization is ubiquitous. More and more of the tasks that used to belong to the "international" team are being done by the generalist engineers. So where will the next advance in i18n sophistication and l10n productivity come from? One important element is to enlist generalist software developers. Change "international" from an inscrutable mystery to a discipline like databases, performance analysis, user experience, or testing: complex and important, but with an essential summary which every developer should know. Teach them the essentials of "international". Help them avoid making innocent mistakes that obstruct i18n later. Show them how they can prepare more for l10n, without having to become "international specialists" and attend IUCs.

This talk is a "train the trainer", by i18n specialists for i18n specialists who will train the non-specialists. It goes through the themes the author has found effective in reaching non-specialist developers. Bring your own ideas for what non-specialists should know. You will leave the talk with a CC-licensed slide deck which you can adapt for teaching your own local non-specialist software developers about "international".

Presenter:

Ken Lunde
Senior Computer Scientist 2, CJKV Type Development, Adobe

Track 3 - Pan-CJK Font Development Techniques, Tips, Tricks & Pitfalls

Picking up where his two similarly-entitled IUC38 presentations left off, Ken Lunde dives deeper into Pan-CJK font development, at a more general level, revealing some of the processes and techniques that were proven to work well for Source Han Sans and Noto Sans CJK development, and detailing some of the pitfalls that were encountered along the way. Genuine Pan-CJK fonts are necessarily complex, mainly due to the enormous number of glyphs that are required, but also because of the requirement to integrate different and sometimes conflicting language- or region-specific conventions. As this somewhat technical presentation will demonstrate, there are proven techniques for handling these characteristics in a way that

makes the resulting installable font resources more user-friendly, which is intended to help font developers avoid some of the hurdles that were eventually overcome while developing the world's first Pan-CJK typeface families.

13:50-14:40 | **SESSION 11**

Presenters:

Sharon Correll
Software Engineer, SIL International

Lorna Priest Evans
Script Technologist, SIL International

Track 1 - New Developments in ScriptSource

An increasing number of people are using ScriptSource as their "go-to" place for information on the world's scripts and writing systems. ScriptSource is a web site for documenting forms of writing around the world and promoting collaboration in development of script-related resources such as fonts and keyboards. It contains an extensive framework for describing and linking languages, scripts, characters, and writing systems, based on data from Unicode, ISO 639-3, the CLDR, and the Ethnologue. The site also includes more general script-related topics such as Unicode, OpenType, locales, and web fonts.

Since ScriptSource first went public in 2011, there have been a number of new features added to the system. One of these is the inclusion of more extensive character exemplar data from the CLDR and SIL's own locale data repository. It is also possible for users to contribute lists of characters for a language they are familiar with. Another new feature is the ability to document how characters are used to represent the phonemes of a language, and investigate how a given phoneme is represented in different languages and scripts. ScriptSource includes an increasing number of links to other language-related sites, such as OLAC and the Endangered Languages project. And since the system now uses web fonts, contributors can be assured that text requiring a special font will be displayed properly.

Presenter:

Nova Patch
Lead Engineer, International Search, Shutterstock

Track 2 - Hello, My Name is ____.

Our personal identity is core to how we perceive ourselves and wish to be seen. All too often, however, applications, databases, and user interfaces are not designed to fully support the diversity of personal, cultural, and gender identities expressed throughout the world. This talk will demonstrate ways to build applications that respect users' identities instead of limiting them.

Topics will include:

- Input, validation, storage, and display of personal names
- Unicode usernames and solutions to security concerns
- Input and use of gender and pronouns
- Internationalization and localization considerations

The intended audience includes programmers, UX designers, and QA testers. Together we can build inclusive software that supports diverse identities.

Presenter:

Alolita Sharma
Senior Engineering Manager, Twitter

Track 3 - Having Fun with Twemoji

Twitter released an open source set of 872 beautiful emojis aptly named the 'Twemoji' set. This set can be used in tweets as well as embedded in web apps. Twemoji is compliant with Unicode 7.0. This talk will explore the Twemoji character set, code to embed in web pages and apps as well as examples of its usage.

14:50 – 15:10 - Afternoon Refreshments

15:10 - 16:00 | **SESSION 12**

Presenter:

Yona Chen
Senior Project Manager, Huawei

Track 1 - Remote L10N Test Solution

Linguistic Testing (LT) is a good way to ensure the quality of localized products. Usually we have two modes, onsite or screenshots. If onsite, testers come to office and test devices as per test cases, but this requires a lot of devices and PCs if you have tens of languages at the same time. Also you need to pay a lot for flights, hotels, and traffic. If you just send screenshots to testers, all you have to do is book their schedule, send all the screenshots to them, and wait for bug reports. But you may worry about the effect because they don't even know what product they test, how it is used, what will popup if you click a menu or button, is there any function issues for the localized version. Remote testing center allows testers to access through Internet/Intranet and test devices according to the test case, and report bugs online. It also provides API to connect with Company string management system, so all the string related bugs can be fixed automatically, thus saving a lot of efforts and improving efficiency and quality.

Presenters:

Erwin Hom
Internationalization Engineer, PayPal, Inc.

Michael McKenna
I18n Product Owner,

Track 2 - Continuing Adventures Going Door-to-Door Around the World

Supporting Postal Addresses.

If your application deals with postal addresses for multiple countries, this talk will highlight the challenges (and solutions) in supporting them in an internationalized application.

In this talk, we'll present:

- Variations of Postal Address Formats

PayPal, Inc.

- Complex Address Formats in UK, Brazil, Australia, Turkey, Japan, China, and others
- The Street Complement component in the Indian Address Format
- Reading an Address as a Native speaking user versus reading it as a Non-Native (for example, English versus Chinese readers)
- Providing for local and international formats to help in cross-border trade
- Dynamic handing of postal address and personal name formats in the UI - address entry forms
- How much do you translate for entry pull-downs?
- Validating address data in the app
- Devising an XML layout syntax based off the open source CLDR and LDML to put name and address layout meta data in a machine-readable form
- A look at open source JavaScript libraries for handling postal address scenarios.

This talk has been updated and improved based on community feedback, more involvement with the open source community, and experience actually working with customers on every continent.

Track 3 - Emoji Search

Presenter:

Nova Patch
Lead Engineer, International Search, Shutterstock

Expanding your existing search software to support emoji queries is easy to implement and provides a fun way for users to navigate and find content, especially on mobile devices. This presentation will demonstrate how Shutterstock implemented emoji search to discover over 50 million images, videos, and audio tracks.

Topics will include:

- Parsing search queries containing emoji
- Expanding your existing search software to support emoji, including Solr- and Elasticsearch-based solutions
- Applying UTR #51: Unicode Emoji and the Emoji Annotation Charts
- Tailoring the use of emoji to best fit your domain
- Localizing emoji queries to search for content in different languages

16:10 - 17:00 SESSION 13

Track 1 - Mobile I18n Testing Toolbox: 2015 Updates

Presenter:

Katsuhiko Momoi
Staff Test Engineer, Google, Inc.

Google has been working on mobile testing tools and approaches to effficiently conduct i18n testing. In IUC 38, I reported on the initial set of tools/APIs from this effort and the plan for open sourcing them. In this presentation I report on the additional new tools that have been created during the past year and provide updates on existing tools and open sourcing status. Our new tools efforts include creation of CLDR data for pseudolocales (for the LTR and RTL pseudolocales) and Mock IME apps/tests for Chrome brower and iOS that check keyboard compatibility against a given input box in an app. I will discuss both existing and new tools with examples and illustrate what sort of issues they are designed to catch. Through these tools and best practice recommendations, we are able to define basic level i18n test coverage for mobile apps. In this regard I will discuss our ongoing efforts to provide concrete steps that product test teams can take to ensure a base level i18n sanity check.

Track 2 - Performance Matters - A New Algorithmic Approach for Fast Unicode Regular Expression Processing

Presenter:

Robert Cameron
Professor, Simon Fraser University

While traditional regular expression processors are fundamentally sequential in nature, processing one input code unit at a time and possibly even backtracking, a new algorithmic approach offers dramatic performance benefits by processing large blocks of input in parallel. The method has been fully implemented in icGrep 1.0 - a fast modern grep implementation with full support for the Unicode Level 1 requirements of Unicode Technical Standard 18 as well as substantial support for legacy Posix and Perl-compatible features. Development of methods for high-perfomance implementation of Unicode Level 2 requirements is well underway. In this presentation, we briefly introduce the underlying method of bitwise data parallelism in regular expression matching, and then move on to discuss the implications of improved performance including the possibility of new Unicode processing tools incorporating higher-level regular expression concepts.

Track 3 - Unicode Emoji: How Do We Standardize that Je ne Sais Quoi???

Presenters:

Maggie Ronan

Release Manager, Indiegogo

Liat Berdugo
Artist & Curator

While it may be tempting to characterize emoji as trendy or childish, their massive popularity demonstrates that they are filling a human need previously unmet by our earlier methods of text-based communication. Millions of people worldwide are making use of these visual elements that evade strict definition to express themselves in ways previously unimagined. As the Unicode Consortium continues on its undertaking of representing commonly used emoji in its standard library, we ask, what are the implications of standardizing a linguistic system in which ambiguity and elasticity are core features?

Nearly as soon as online comment boards and chat rooms came into existence, internet users began lamenting the inability to use them to accurately express nuanced feelings through the short form written word. We see this effect compounded in text messaging, an endeavor often undertaken in a contextless and expeditious manner-a few simple words shot off with the intention of clarifying plans can be received as uncaring and cold. Enter the emoji: an expressive, interested friend (that can be typed just as easily as the next standard character) ?! With a quick character selection from the library available to us on our

phones, suddenly a curt and business-like one-liner becomes softened and familiar ?. In a 2014 New York Magazine piece, journalist Adam Sternbergh observes that the "elasticity of meaning is a large part of the appeal and, perhaps, the genius of emoji. These seemingly infantile cartoons are instantly recognizable, which makes them understandable even across linguistic barriers. Yet the implications of emoji-their secret meanings-are constantly in flux."

There is a seemingly endless flood of controversies drummed up by bloggers and tech journalists over the various ways to interpret specific emoji in the vein of one of the hottest questions of September 2014 (e.g. just what *did* the rapper Drake intend to convey by emblazoning his forearm with a U+1F64F "Person With Folded Hands" (?) tattoo?) In other realms, artists are seeking-sometimes cheekily-to reach across language divides and unite all readers with books written completely in Emoji (the US Library of Congress accepted its first emoji book, Emoji Dick, in 2013).

We know the dilemmas that emojis present to Unicode quite well. As Unicode Consortium Vice President Peter Constable said, "With most text, you don't have things being invented left, right, and center. The letters of English are the letters of English. We don't have people inventing new letters of English every day." A unique challenge presented by emoji is that there is no end to the symbols that could be proposed. Emoji standardization also faces a diversity challenge-both along axes of race and sexuality. But the more conceptual dilemma to address is: what does it mean to standardize this linguistic token which, by its very nature, resists linguistic standardization?

As cultural critics and theorists-as well as emoji connoisseurs-who have devoted significant portions of our lives to exploring the relationship between technology and society, we are uniquely positioned to examine these questions of how technological decisions made in the effort to standardize can impact the meaning of a linguistic system itself. Our day-to-day lives as an artist/writer/curator/professor focused in new media (Liat) and culture-enthralled tech community enthusiast (Maggie) call us to ponder these questions on the regular: we use emerging technologies like Slack to create community-specific dialects with custom emoji; we engage with friends who found and operate ? ?(pronounced "disk cactus", the first company with a native emoji name!), the Oakland-based art and technology studio, and makers of the Emoji Keyboard; and we express ambiguous emotions and complex ideas with those we love the most in the world using these standardized glyphs.

As the Unicode Consortium continues its bold efforts in standardization of emojis, it is imperative that we take a step back and examine the social landscape that surrounds their use. We find that much of not only the delight, but the utility, of emoji comes from their ability to convey ambiguous meaning, and it is our hope that as we continue to set standards to ease the use of emoji across platforms and cultures, we keep in mind the beauty that lies in the ambiguous.

Program is subject to change.