Internationalization & Unicode Co
Local. International. Global. Unicode. 3
OCTOBER 22-24, 2012 • SANTA CLARA, CA USA

Program - Session Descriptions

Monday, October 22, 2012

08:30-10:00 **MORNING TUTORIALS**

Presenter:

Richard Ishida
Internationalization Activity Lead, W3C

Track 1: An Introduction to Writing Systems & Unicode

The 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 also 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:

Addison Phillips
Globalization Architect, Lab126 (Amazon)

Track 2: Internationalization: An Introduction, Part I: Characters and Character Encodings

What is internationalization? What do developers, product managers, or quality engineers need to know about it? How does a software development organization incorporate internationalization into the design, implementation, and delivery of an application?

This tutorial track provides an introduction to the topics of internationalization, localization and globalization. Attendees will understand the overall concepts and approach necessary to analyze a product for internationalization issues, develop a design or approach, and deliver a global-ready solution. The focus is on architectural approaches and general concepts, but will include specific examples and exercises.

Part I focuses on characters, character encodings, and the basics of Unicode.

Presenter:

Track 3: Tutorial Web Internationalization - Standards and Best Practices

Tex Texin
Chief Globalization Architect, Xencraft

This tutorial is an introduction to internationalization on the World Wide Web. The audience will learn about the standards that provide for global interoperability and come away with an understanding of how to work with multilingual data on the Web. Character representation and the Unicode-based Reference Processing Model are described in detail. HTML, including HTML5, XHTML, XML (eXtensible Markup Language; for general markup), and CSS (Cascading Style Sheets; for styling information) are given particular emphasis. The tutorial addresses language identification and selection, character encoding models and negotiation, text presentation features, and more. The design and implementation of multilingual Web sites and localization considerations are also introduced.

10:00-10:15 - Morning Refreshments

Presenter:

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

Richard Ishida
Internationalization Activity Lead, W3C

The 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 also 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:

Track 2: Internationalization: An Introduction (Part II, Writing Global Ready Code)

Addison Phillips
Globalization Architect, Lab126 (Amazon)

Part II focuses on preparing for the localization (translation) of user interfaces; making applications "locale-aware", including format and display differences; as well as approaches to delivering multi-lingual and multi-locale software or content.

Presenter:

Track 3: Using ICU Workshop

Steven R. Loomis
Software Engineer, IBM

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.

12:00-13:00 - LUNCH

13:00-14:30 **AFTERNOON TUTORIALS**

Presenters:

Craig R. Cummings
International Engineering Leader, Zynga, Inc.

Michael McKenna
International Engineering Leader, Zynga, Inc.

Track 1 - Unicode - A Grand Tour

This tutorial covers the next level of detail of what Unicode is, and how it is used in the real world. The modules of the tutorial will cover: The Unicode standard - what are the "Guiding Lights", or design principles behind Unicode? A tour of Unicode's structure, encoding forms, behavior, technical reports, database, and how to use the Unicode Standard. Implementation according to Unicode - a walk through the details of attributes, compatibility, non-spacing characters, directionality, normalization, graphemes, complex scripts, surrogates, collation, regular expressions and other aspects according to the Unicode Standard and associated Technical Reports.

Unicode and the Real World - an overview of International Components for Unicode (ICU) and implementations supporting Unicode in web servers, application servers, browsers, C/C++, Java, PHP, SQL, and various operating systems. On-going programs - how Unicode is evolving to support more minority scripts, languages, and help solve linguistic processing issues.

This tutorial will be updated to reflect latest changes for Unicode with updates for 6.1 and possibly 6.2.

Presenter:

Pravin Dinkar Satpute
Senior Software Engineer, Red Hat

Track 2 - Developing an OpenType font for complex scripts using Fontforge

In recent versions Unicode has added many new South-Asian scripts. Free open-source OpenType (OT) fonts for these newly added scripts are not yet available, and user communities are still waiting for fonts for these scripts.

The major reason for this is developing a OT font for a complex script is both time consuming and difficult task. Expertise is required from different domains like Design/Calligraphy, Linguistics and Technical.
This tutorial will explain the steps by step process to create an OT font for the Devanagari script using the Fontforge application and cover the concepts involved in OT font development.

After attending this tutorial one will understand how to create OT fonts for complex script and in the coming years we hope to see more contributions of OT fonts for newly added Unicode scripts.

Body:

Indian scripts are complex. The keys we press and what we see on the display are often not the - lots of processing is involved between. Developing a font for such a complex script is further complicated since expertise is required from following different domains.

Linguistic:

Knowledge regarding script and properties of its characters is must, one can't start font development without this knowledge. Complex script involves lots of characters as well some ligatures. Ligature gets formed by combining basic characters.

Design/Calligraphic:

For designing a glyph shape of a particular script one should have calligraphic knowledge. Once we design a glyph, one need to digitize it with the a font editor.

Technical:

-Understanding of various element involved in rendering
- Understanding of the Open type specification.
- Layout engine (Uniscribe, Pango, QT/harfbuzz and ICU)
- OpenType tables (GDEF, GSUB and GPOS)

In this tutorial we will see all these points with a demonstration of the Lohit Devanagari font.

References:

http://unicode.org/
http://www.microsoft.com/typography/default.mspx
http://fontforge.sourceforge.net/
https://fedorahosted.org/lohit

Presenter:

Martin J. Dürst
Professor, Aoyama Gakuin University

Track 3 - 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 using the programming language 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 works in Ruby and how to make the best use of it both in thow-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 preferably uses UTF-8, 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 show discuss both the bacis framework as will 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 Web applications. Experience with Ruby and/or Ruby on Rails is a plus, but is not a precondition for attending.

14:30-14:45 - Afternoon Refreshments

15:45-17:45 **AFTERNOON TUTORIALS**

Presenters:

Michael McKenna
International Engineering Leader, Zynga, Inc.

Craig R. Cummings
Globalization Center of Excellence, Rearden Commerce

Track 1 - Unicode - A Grand Tour (Cont'd.)

This tutorial covers the next level of detail of what Unicode is, and how it is used in the real world. The modules of the tutorial will cover: The Unicode standard - what are the "Guiding Lights", or design principles behind Unicode? A tour of Unicode's structure, encoding forms, behavior, technical reports, database, and how to use the Unicode Standard. Implementation according to Unicode - a walk through the details of attributes, compatibility, non-spacing characters, directionality, normalization, graphemes, complex scripts, surrogates, collation, regular expressions and other aspects according to the Unicode Standard and associated Technical Reports.

Unicode and the Real World - an overview of International Components for Unicode (ICU) and implementations supporting Unicode in web servers, application servers, browsers, C/C++, Java, PHP, SQL, and various operating systems. On-going programs - how Unicode is evolving to support more minority scripts, languages, and help solve linguistic processing issues.

This tutorial will be updated to reflect latest changes for Unicode with updates for 6.1 and possibly 6.2.

Presenter:

Edwin Hoogerbeets
CTO, JEDLSoft

Track 2 - I18N in Javascript with ILib

It used to be that you would do i18n on your app server and ship the already-formatted locale-sensitive text to the browser. Now with the tremendous rise in popularity of AJAX and the proliferation of web services, many of which are not provided by your own app server, it is no longer possible to avoid doing i18n on the browser side. The standard Javascript libraries available in most browsers are pretty meager when it comes to i18n classes, but will improve significantly when the proposed ECMAscript internationalization proposal is accepted and implemented everywhere. However, the ECMAscript proposal only includes a few i18n classes, and is not yet a complete set. This tutorial shows how to use the new open-source library called ILib that is available today to perform a number of i18n functions to create a fully globalized, AJAX-enabled site, including a number of functions that are not part of the ECMAscript proposal.

Presenter:

Michael Kuperstein
Localization Engineer, Intel Corporation

Loïc Dufresne de Virel
Localization Strategist, Intel Corporation

Track 3 - The Road to World-Class Starts with World-Ready

Join us in an interactive tutorial session where we'll guide you through a systematic process, using funny and practical examples of typical every-day internationalization mistakes, to internationalize a sample application. In one afternoon, you might not become an I18N expert, but you'll gain a thorough understanding of the issues and practical knowledge of the tools and processes that are available within the industry to help you develop truly World-Ready applications.

As more and more apps are accessible anywhere and usage models move from desktop-based to cloud-based, the need to release World-Ready software is clear. Applications need to offer the same (or equivalent) user experience to all users, regardless of the language and locale preferences of their device, operating system, or browser. They need to offer the same level of functionality regardless of the script and language used to exchange data with the rest of world. It just has to work!

Sadly, often times it doesn't… perhaps not fully, not completely, or not at all. Experience shows that products are often released with internationalization (I18N) defects, or companies only start addressing those defects during localization, because they don't know what to look for during the initial development stages, when the focus is on releasing an English-only product. Down the road, this causes rework, additional costs and delays: things that could be avoided by designing with world-readiness in mind. This tutorial is structured to give you a comprehensive and practical overview of what being world-ready means: definitions, processes, tools, and sample issues. We'll go through the whole nine yards to get you on your way to World-Class!

Presenter:

Amit Gupta
Member Technical Staff, Adobe Systems

Track 1 - Internationalizing Domain Names in Applications (IDNA)

This presentation details internationalized domain names (IDNs) and a mechanism called Internationalizing Domain Names in Applications (IDNA) for handling them in a standard fashion. IDNs use characters drawn from a large repertoire (Unicode), but IDNA allows the non-ASCII characters to be represented using only the ASCII characters already allowed in so- called host names today. This backward-compatible representation is required in existing protocols like DNS, so that IDNs can be introduced with no changes to the existing infrastructure. IDNA is only meant for processing domain names, not free text.

IDNA works by allowing applications to use certain ASCII name labels (beginning with a special prefix) to represent non-ASCII name labels. Lower-layer protocols need not be aware of this; therefore IDNA does not depend on changes to any infrastructure. In particular, IDNA does not depend on any changes to DNS servers, resolvers, or protocol elements, because the ASCII name service provided by the existing DNS is entirely sufficient for IDNA.

Applications can elect to use IDNA in order to support IDN while maintaining interoperability with existing infrastructure. If an application wants to use non-ASCII characters in domain names, IDNA is the only currently-defined option.

Presenter:

Track 2 - Keyboard design for Tavultesoft Keyman and Unicode

Marc Durdin
CEO, Tavultesoft Pty Ltd

This tutorial works through designing an input method that will work across multiple platforms, using Tavultesoft Keyman technology.

Starting with some principles of cross-platform input, we will finish with an input method for the Lao language that works consistently across Windows, Mac OS, iPhone, Android and iPad platforms, and presents itself in each case in a platform-appropriate manner and even takes advantage of platform-specific functionality. The same keyboard can also be targeted for websites.

Some time will be spent discussing character ordering, constraints (preventing invalid combinations), normalization during input, and designing efficient input methods. Phonetic and visual methods will be covered.

An understanding of basic Unicode principles will be helpful but no prior experience of keyboard design is required.

Presenter:

Track 3 - Building multilingual websites in Drupal 7 and Joomla 2.5

Jim DeLaHunt
Principal, Jim DeLaHunt & Associates

A practical look at the language and locale capabilities of Joomla! 2.5 and Drupal 7, two leading free software content management systems (CMSs). They let you build more powerful, more international websites faster. We look at: their core internationalisation and locale services; localisation of UI and content. Each platform just had a major release, with advances in internationalisation. You will leave with specific tips for building your own site. We don't assume Joomla or Drupal experience, but do include material for advanced practioners. A good tutorial for web site product managers, web designers, developers, and managers of international web teams.

Tuesday, October 23, 2012

09:00-09:15	***WELCOME & OPENING REMARKS***
09:15-10:00	***KEYNOTE PRESENTATION -*** **"Bit Rot" – A Disaster Waiting to Happen**

Presenter:

Dr. Vinton G. Cerf
Vice President and Chief Internet Evangelist, Google

Cerf will discuss the problem of curating digital content on the order of centuries. Unicode has a role to play although there are very complex issues relating to format and structure of digital objects, interpretation of content, intellectual property management, perhaps even patents and other legal framework questions. The problems are both technical and legal.

10:00-10:30 - Morning Refreshments

10:30-11:20	**SESSION 1**

Presenters:

Track 1 - A Platform for “Screen To Text"

Yong Zhang
QE Developer, Adobe China R&D

Text translation is commonly used in daily life. However, most translation applications require text to be typed in manually, which is hard for a user who wants to translate text of an unknown language. Meanwhile, common text translation can't

Center

Jing Lai
QE Developer, Adobe China R&D Center

Shiyao Bao
QE Developer, Adobe China R&D Center

handle non-editable text, such as text in images, photos and so on. Thus, a solution to pick up visible text is needed. This proposal introduces a solution and a platform for "Screen-to-Text". User may select an area on screen via mouse(on desktop) or finger(on mobile), which is part of a UI, photo, image, or any visible content on screen with text, then the selected area is saved as image and sent to server automatically. Once server receives the image, it runs pre-processing step to improve image quality, and then retrieves text with OCR technology. The recognized text will be handled by various applications in post-processing and return results to client to meet different customer requirements.

The main advantage of this platform is that it doesn't need manual typing, and it can handle all visible text even within images or photos. Compared with other OCR applications, it has the following strengths:

- More powerful - this platform is based on the OCR technology in Adobe Acrobat, which supports more than 40 languages.
- More accurate: the pre-processing makes use of Adobe Photoshop technology, which improves the quality of received images and helps to enhance OCR results.
- More flexible and extensible. This platform can be extended to develop various applications upon and that's why it is a "platform", not an application. For example, a translation application is easily created if combining text reorganization with machine translation SDK. This application helps software engineer with software internationalization and localization. It is also useful in real world. People may take pictures with mobile device of foreign text, and then get the translated result via the application. A demo application for translation is implemented already. Another sample application provides pronunciation functionality, in which text recognition will be combined with "Text to Speech" SDK, so that text can be read out in client. In addition, text recognition can also be implemented in search applications once integrated with search engine API. In conclusion, this platform is flexible and extensible to develop applications not only for text translation, but also for different business goals.

This platform can be integrated with various technologies, such as cloud, machine translation, text to speech, search, social network and so on. It is available on both desktop and mobile platforms. In future, the self-learning mechanism will be introduced into the platform to enhance the performance and accuracy of text recognition.

Presenter:

Track 2 - New HarfBuzz Coming to a Device Near You

Behdad Esfahbod
Software Engineer, Google

HarfBuzz is the Free text shaping engine based on Unicode and OpenType standards. It is a centerpiece of Unicode text rendering on Linux, Android, ChromeOS, and Firefox, among others.

In this talk I will discuss the current development status of the HarfBuzz rewrite (aka harfbuzz-ng), design decisions made, our testing infrastructure, test-driven development model, and future plans.

Presenter:

Track 3 - Globalizing Text Analytics Applications

Sudhakar Pandey
Computer Scientist, Adobe Systems

Text analytics essentially means applying NLP (Natural Language Processing) techniques to extract meaningful and qualitative information from the text document and then use this information for various applications. NLP is field of Computer Science and Linguistics concerned with the interactions between computers and human (natural) languages. Typical Text analytics Applications are Automatic summarization, Machine translation , Auto Tagging, Sentiment Analysis etc. Now a days, most of the software developers do take care of Internationalizing their code to enable the support of multiple languages whenever the need arises, but beside the best the recommended Internationalization coding practices that we do for conventional software desktop and web software's, Text analytics Applications involves additional knowledge and understanding of the languages needs as these apps have lot of language specific low level modules. During the presentation, We would be talking about what these language specific low level modules are and we would understand each of them with examples for different languages.

Whenever one starts developing Text analytics Application, say for English(US). Without realizing , the developer implements lot of modules and rules which would be only applicable for English language, However those rules and modules would require drastic changes for supporting other languages. Let me take an example of a very common Text analytics Application e.g. "Search Application". First of all, we need to process a lot of multi-lingual documents to extract keywords, index them and store them in a database. So that when the user provides the search query later, it retrieves the relevant documents from the database. This whole process involves lot of modules which are language specific. We would first need a language detector module to find out the language of the document , then we need a Paragraph Segmentation and tokenization module to split the document into paragraphs, sentences and finally into words. If the document is for languages like German, Dutch which supports compound words , we also need to decompound the tokenize words. Similarly we would require lot of other language specific modules like Stemmer, Chunker etc.

We would be touching upon these language specific needs of Text analytics Applications, we would understand each of them with examples for different languages and how we should take care of Globalizing our Text analytics Applications by provisioning for these modules at appropriate places to avoid the hassle later.

11:30-12:20 SESSION 2

Presenter:

Marc Durdin
CEO, Tavultesoft Pty Ltd

Track 1 - From Typewriter to Touch: Multi Platform Keyboards -- Challenges and Illustrations

This presentation is a walk through the world of modern keyboard input methods, comparing desktop and laptop hardware keyboards with tablet and mobile touch-screen keyboards, and looking at the challenges presented when entering characters from the thousands of languages and many different scripts supported by Unicode. Complexities with text input, selection, insertion and deletion, directionality, phonetic-style input, autocorrection and language tagging will be discussed, with examples that illustrate some of the less obvious problems that application designers and input method developers may encounter. Some time will be spent examining the common basis underlying traditional physical keyboards and touch-based input methods, and how the platforms diverge.

Presenter:

Guy Smith-Ferrier
Internationalization Consultant, Capella Software Ltd.

Track 2 - How To Achieve World(-Ready) Domination In ASP.NET MVC

So you've written your ASP.NET MVC application and you want it to work in another language? Then this session is for you. World-Readiness is all of the work that a developer needs to do to globalize an application and make it localizable (i.e. capable of being localized). In this session we will cover localizing HTML and HTML Helpers, localizing and globalizing Data Annotations, localizing and globalizing JavaScript and localizing URLs. No previous experience of ASP.NET localization is required.

Presenter:

Luke Swartz
Product Manager,

Track 3 - Innovations in Internationalization at Google

This talk will cover a range of internationalization challenges that Google has encountered and overcome in the past year. Among the topics are i18n of personal names, how to better serve multilingual users, expanding CLDR and ICU (and Google products) to more languages, improvements in ICU MessageFormat Syntax, Bidi wrapping, web app i18n, and input technologies.

Google Inc.

Mark Davis
Sr. Internationalization Architect, Google Inc.

12:30-13:30 - LUNCH

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

Presenter:

John O'Conner
Globalization Architect, Adobe Systems

Track 1 - Determining and Prioritizing Language Signals in Web Applications

Traditional desktop software often offers installers that provide customers the opportunity to select and use a preferred user interface language. Additionally, the preferred language is typically set in the operating system itself, providing desktop applications with a strong language hint. However, browser based applications don't have a single strong language signal. Instead, web applications can receive multiple and conflicting language signals for a customer's preferred language. This session explores the multiple UI language preference signals that are available to web applications and suggests a prioritization for handling them.

Presenter:

Markus Scherer
Unicode Software Engineer, Google

Peter Edberg
Senior Software Engineer, Apple Inc.

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 49, including the latest support for Unicode 6.1 and CLDR 2.1, date/time formatting & parsing improvements, and other changes (see http://site.icu-project.org/download/49). The presentation will also touch on ICU's planned direction for future releases.

Presenter:

Daniel Yacob
Director, The Ge'ez Frontier Foundation

Track 3 - Computer Typography Challenges of the Ethiopic Zaima Notation Practice

The Ethiopian Orthodox "Yaredic" Zaima annotation system is a roughly 1400 year old quasi-neumic ekphonetic practice for recording and recitation of liturgical chants. Quite possibly the most complex system of musicological notation, the training demands of the practice have made it the exclusive subdomain of specialists within the Ethiopian Orthodox clergy.

Since its divinely inspired inception by Ethiopia's 6th century Saint Yared, Zaima annotation remains a calligraphic practice to this very day. With the introduction of the basic "tonal marks" beginning from the Unicode 4.1 standard, the practice has been primed to leap across two millennia into the era of computer typography. But is the best of 21st century typesetting technology ready to receive it?

The presentation will review the history and role of the calligraphic tradition within liturgy, both graphical and lexical components, logical rules and model. Conventions are compared and contrasted with the superficially similar Ruby convention for annotation of Asian scripts. Typesetting challenges will be presented throughout and the limitations found in both software capabilities and the Unicode Standard will also be illuminated.

The presentation represents a nearly final milestone in a multi-year effort of The Ge'ez Frontier Foundation to define the software requirements needed to support the modern typography of the practice with the same precision as the Yaredic calligraphy. The topic will be of particular interest to software engineers with a focus on word processors and typesetting technology, typeface developers and ethno-musicologists.

14:30-15:20 SESSION 4

Presenter:

Track 1 - Internationalization Requirements for Indic Language Layout in CSS

Swaran Lata
Director, HoD, TDIL Programme and Country Manager W3C India

Somnath Chandra
Joint Director, Dept of Electronics & Information Technology, Govt of India

Cascading Style Sheet (CSS) is one of the most important building blocks for design of web page and being used in present and next generation web mark-up languages such as HTML 5.0 and e-publishing framework. CSS enables web designer to incorporate the cultural and linguistic requirements while designing multilingual web sites. The present CSS rules while applied for designing the web pages in Indic languages does not render the complex Indic characters properly and also depend on implementation strategy of different browsers. The complex mapping relations between Indian script and languages demands separate linguistic rules to be incorporated in the CSS standard. The paper describes the nuances of Indic languages while implementation of CSS and possible approach to overcome the problem using possible modifications of Unicode algorithms.

The present study is based applications of CSS rules in eight most important Indian Languages namely Hindi, Bengali , Marathi, Gujarati, Tamil, Telugu , Malayalam and Kannda. The paper also presents automata based definition Indic Akshara (character) to the address the issues of Unicode line breaking and segmentation algorithms, which will enable rules for various CSS components such as First Letter, Vertical and Horizontal arrangement of Characters, Numbering, Underlining of Characters, Line breaking & Segmentation , Letter Spacing , Word-wrap and indention.

Presenter:

Track 2 - Localizing Windows 8 Metro Apps Using The Multilingual App

Guy Smith-Ferrier
Internationalization Consultant, Capella Software Ltd.

The Multilingual App Toolkit is a free Visual Studio add-in from Microsoft. It translates, packages and unpackages resources for round tripping with localizers. This session illustrates its use for localizing Windows 8 Metro applications and shows how developers and localizers can work together to build localized applications.

Presenter:

Track 3 - Arabic Typography

Thomas Milo
Partner, DecoType

Coming soon...

15:20-16:00 - Afternoon Refreshments

SESSION 5

Presenter:

Track 1 - Go Global But Not Alone – Multilingual SEO

Anubhav Jain
Computer Scientist, Adobe Systems

English is not the native (first) language for over 70% of Internet users. This means that there are 900 million users whose search behavior is very different from that of native English speakers. If your business is (or could go) global, this session will give you more insight and skills to create multilingual websites with effective international search engine optimization (SEO).

This session is ideal for those familiar with the principles of SEO, and are also looking to develop or improve upon global strategies for their business, and learn about current developments in global SEO.

The internet continues to grow, and has become the default point of call for businesses and individuals searching for goods, services, or information. For businesses aspiring to have a competitive advantage, a multilingual website is one of the most high impact means of expanding a client base and securing greater sales volumes. Multilingual websites will continue to become a necessity for such businesses and organizations. However, before a business moves forward with its globalization plans, it must consider what SEO means in the international arena. Businesses know that customizing their websites to suit the linguistic needs of regional markets can result in significantly increased Web traffic, as well as increased revenue per order and improved international brand recognition. However, merely translating a website into another language is not enough to take advantage of the potential for expansion in foreign markets. International SEO (ISEO) is a key ingredient to such success. This session will share the best practices in International SEO, clarify certain myths, and will be centered on the following -
To begin with, a case study of top e-commerce companies that have gone global, and how following ISEO practices have helped them in getting ahead of the competition
Know more about multilingual search engines

- Google is not the only search engine
- How a search engine determines the geographic intent of the searcher
- How search engines use NLP (Natural Language Processing) for language detection and practices in SEO
- Unicode and SEO
- How friendly are Unicode URLs for SEO?
- Character set issues with ISEO
- Domain name strategy for ISEO - Is having one domain better?
- SEO penalty - What does it mean in the ISEO universe, and how to avoid/overcome it.
- How does hosting location for a site affect its visibility?
- Myth - Auto-redirect based on perceived user language provides a good user experience.
- Do our words and images convey the right message in every language? How to leverage them for ISEO.
- Design your site structure and content to drive better ISEO. Designing landing pages and websites for international users.
- Impact of translation quality on ISEO. Is machine translation SEO friendly?
- How Social Media Optimization (SMO) capitalizes on the large volume of traffic social networking websites bring, to enhance ISEO.
- How to create, implement, and manage multilingual SEO campaigns.
- Best practices for international pay-per-click engines (PPC), and examples covering a variety of international PPC engines.
- Having an international strategy in place is not enough. You must monitor and measure its performance regularly, tweak it based on feedback and geographies.

Presenter:

Markus Scherer
Unicode Software Engineer, Google

Mark Davis
Sr. Internationalization Architect, Google Inc.

Track 2 - Plural & Gender & More in Translated Messages

"There are 1 file(s)." / "Alice added 1 people to his mailing list." - User-facing messages with placeholders for numbers and strings are common technology. These require the placeholders and text to be reorderable to account for grammar of different languages. However, the common technology does not solve the problem of plural and personal gender in placeholders. That is, depending on the language and the placeholder values, the surrounding text often needs to change, as illustrated by the examples above.

ICU has been improving on the Java formatting framework, adding support for such message variants in both its C++ and Java versions. More recently, we have added support for ordinal-number variants and formatting of lists of items. This session explains the challenges, approaches, and new functions and capabilities.

Presenter:

Track 3 - Polyglots in the Mist

Michael Erard
Linguist

Author Michael Erard talks about his latest non-fiction book, Babel No More: The Search for the World's Most Extraordinary Language Learners (Free Press, 2012), in which he answers the question: What is the upper limit of the ability to use, learn, and remember languages? In the book, he travels the world to locate and make sense of "hyperpolyglots," gifted language learners who are also massive language accumulators, including famous historical figures and contemporary language learners who are responding to new environments, resources, and emerging forms of multilingualism, many of them online. The Economist wrote that "[Erard] approaches his topic with both wonder and a healthy dash of scepticism...feeling his way through his story as a thoughtful observer, rather than banging about like an academic with a theory to defend or a pitchman with a technique to sell...fascinating."

17:00-17:50 SESSION 6

Presenters:

Track 1 - Bringing Multilingual PDFs to The Open Web

Adil Allawi
Director, Diwan Software

Mozilla Foundation's PDF.js project has the humble aim to render Adobe's Portable Document Format (PDF) using only web standards. I became involved with the project a year ago. Despise many people telling me that HTML and Javascript were not ready to render a complex format like PDF, this year PDF.js will become a standard feature of the Firefox web browser. This talk will cover the problems we faced converting the variety of multilingual fonts and encodings in the PDF standard to web standard formats.

As a standard that is as old as Unicode itself, PDF, in its own way, embodies the multitude of legacy encodings, font hacks and bad practices that predated the widespread adoption of Unicode. PDF provides a solution for keeping printed documents in an electronic format that could be used in the future, and it has been extremely successful. However, for many non-Roman languages the PDF may be no more than an image of the printed page. If the font had not been carefully encoded or named, and the application was not cooperative, extracting the text from a PDF can be challenging to say the least.

This paper will look at how multilingual text and fonts have been encoded in PDFs and the methods the PDF.js project used to convert these over to modern web standards.

There is still a long way to go and many documents can still not be converted to searchable text. I will especially cover issues of converting complex legacy non-Roman languages, like Arabic and Chinese, from PDF to Unicode and web fonts. Looking at solutions that exist and the work that needs to be done.

I will go on to talk about the problems with current web standards for rendering complex documents and how these standards need to change in the future.

Presenter:

Track 2 - Internationalizing the Core of JavaScript

Norbert Lindenberg
Internationalization Consultant, Lindenberg Software LLC

Nebojša Ćirić

ECMAScript, the standard at the core of JavaScript, is currently seeing significant internationalization improvements. In the language specification, support for Unicode supplementary characters is being mandated, and time zone support is improved. The big news, however, is the new ECMAScript Internationalization API, which provides collation, number formatting, and date and time formatting. This talk will introduce the specification changes and new APIs, and show off the latest implementations in browser and server systems.

Software Engineer, Google, Inc.

Co-presenting are Neboja Ciric and Jungshik Shin of Google, who presented initial ideas for the Internationalization API at the Internationalization and Unicode Conference in 2010, and Suresh Jayabalan of Microsoft.

Jungshik Shin
Software Engineer, Google, Inc.

Suresh Jayabalan
Program Manager, Microsoft

Presenters:

Track 3 - Designing a Panlingual Dictionary

Jonathan Pool
Director of the PanLex Project, The Long Now Foundation

Susan M. Colowick
Research Associate in the PanLex projects, The Long Now Foundation

Laura Welcher
Director of Operations and Director of The Rosetta Project, The Long Now Foundation

PanLex, a project of The Long Now Foundation, seeks to become a panlingual dictionary, integrating every known lexical translation into a publicly accessible database. About half a billion translations among about 18 million lexemes in over 6,000 languages have been documented so far. PanLex would have been impractical without many of the sources used by this project being Unicode-compliant, and without the ability to recode other sources' legacy data into Unicode. But normalization and standardization remain challenges far beyond Unicode compliance. Designing a panlingual dictionary has also entailed principled decisions about character repertoires, character equivalence, transliteration, extra-Unicode scripts, language-variety identification, lemmatic forms, lexemic status, synonymy, punctuation, letter case, and grammatical classification. All these issues involve controversy and trade-offs. I describe normalization and standardization decisions made so far in the PanLex project, hoping that audience members will help us tune these decisions for the applications into which they envision incorporating PanLex.

18:00-19:00 - IUC36 CONFERENCE RECEPTION

Wednesday, October 24, 2012

09:00-09:50	SESSION 7

Presenter:

Track 1 - Internationalizing the Kindle Paperwhite

Addison Phillips
Globalization Architect, Lab126 (Amazon)

Presenter:

Track 2 - Unicode and Legacy Representations of Emoji

David Yonge-Mallo
Software Engineer, Google

Emoji were included in Unicode only relatively recently, but software developers have been using emoji or emoji-like characters for much longer. Many applications have been written which support emoji by using either an encoding other than Unicode or by assigning emoji characters to a Unicode PUA (private use area) code range.

Now that a set of emojis have been made a part of the Unicode Standard, software will have to be modified to use the new code points, while maintaining support for legacy data and compatibility with older versions. In this talk, we outline some of the challenges involved in converting an application from a proprietary system for encoding emoji to using Unicode. We draw upon experiences from Gmail and other products.

Presenters:

Michael S. Kaplan
Program Manager, Microsoft

Track 3 - Show Me The Money!

It has always been a rule at Microsoft that locale data is not changed for prior versions or in service packs for the current version. And lke all rules, there is one major exception: currency values! I'll be taking you on an exciting journey of the many updates for the Euro, the Indian Rupee, the Turkish Lira, and other times that Microsoft has shown its customers the money....

10:00-10:50 SESSION 8

Presenter:

Peter K. Edberg
Senior Software Engineer, Apple

Track 1 - What's New in CLDR?

The Unicode Consortium's Common Locale Data Repository project (CLDR) defines LDML (Locale Data Markup Language) and uses it to organize and provide the most extensive open repository of locale data, with data collected primarily via the web-based Survey Tool. This session provides a brief overview of CLDR, then focuses on recent and forthcoming enhancements including the new BCP47 -t- and -u- extensions, ordinal categories (1st, 2nd, …), context-dependent capitalization, collation reordering (e.g. Cyrillic before Latin), multiple numbering systems for a locale, abbreviated numbers (e.g. "1.2 B"), Chinese lunar calendar data, keyboard layout data, and a completely revamped Survey Tool (which makes it much easier to enter data). Presenters: Mark Davis (Google, CLDR TC chair), Peter Edberg (Apple, CLDR TC member)

Presenters:

Linus Toshihiro Tanaka
Senior Manager, Localization Engineering, Yahoo!, Inc.

Track 2 - Supplementary Characters Revisited -- How Yahoo! Handles World's Characters

Unicode includes more than 110,000 characters. Long time ago it had been a challenge to handle such wide range of characters especially when the characters are outside Basic Multilingual Plane (BMP) which was the original range of Unicode. Characters outside BMP are called supplementary characters, and in the past their use had been somewhat limited to particular areas of the world. However, with some recent systems, everyone can now use supplementary characters, not only for particular languages but with any language. To handle supplementary characters, technologies got improved and many issues were resolved. I explain those issues, how they are addressed at Yahoo!, and what still need to be done.

Unicode has 3 encoding forms, UTF-32, UTF-16 and UTF-8. For supplementary characters, all these 3 encoding forms get 32-bit values. Therefore, for supplementary characters, an important step is to handle 32-bit values. For characters within BMP, UTF-32 still gets 32-bit values, but UTF-16 gets 16-bit values and UTF-8 gets 8-bit, 16-bit or 24-bit values. As a result, it is necessary to go beyond 16-bit values for UTF-16, and beyond 24-bit values for UTF-8, in order to handle supplementary characters.

Yahoo! uses multiple Database Management Systems (DBMS), many programming languages on multiple platforms, and variety of technology stacks. We need to make sure that each of them, and various combinations of them, can handle supplementary characters. This effort becomes more and more important, and it should help not only Yahoo!'s customers and ourselves, but also those who are working on programming languages, development frameworks, and unlimited kinds of software development projects.

Presenter:

Track 3 - Unicode Localization Interoperability: Overview

Uwe Stahlschmidt
Principal Group Manager, Microsoft

Interoperability in exchange of localization data has been a challenge in the localization industry for many years. The Unicode Localization Interoperability Technical Committee (ULI TC) was established in 2011 with the goal to help ensuring interoperable data interchange of critical localization-related assets, including translation memories, segmentation rules, and translation source strings and their translations. ULI is an expert group with representatives from localization service consumers, localization service providers, tools/technology experts, academia and standards organizations who collaborate and advise on interoperable data interchange of critical localization-related assets. The objectives of the TC are to optimize the service time between systems through consistent interpretation and adoption of localization data interchange standards (mature existing standards and data references by gathering requirements for extensions of localization interoperability standards) and reduce cost through best practice guidelines by providing open reference implementation of the extensions and profiles (establish reference implementations or extensions to improve the usefulness of localization interoperability standards). This session will provide an overview of the charter and objectives of the ULI, an overview of the work over the last year, and possible future projects.

10:50-11:10 - Morning Refreshments

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

Presenter:

Track 1 - Round Table Discussion: Locale Data Issues

Mark Davis
Sr. Internationalization Architect, Google Inc.

This panel discussion focuses on issues connected with locale data, particularly use of locale data, locale and language identification (BCP47, etc.), knotty problems for implementers, gathering reliable locale data, UI issues, and so on. The particular topics will be driven by audience and moderator questions.

Presenter:

Track 2 - Normalization of Ideographic Description Sequence

Taichi Kawabata
Researcher, NTT Corporation

Ideographic Description Sequence (IDS) has become crucial information for searching and identifying the CJK Ideographs. However, a single CJK Ideograph could be represented by various IDS patterns, causing ambiguity on IDS-based processing. This presentation describes ambiguity and other difficulties of IDS processing, and propose an algorithm to normalize IDS so that single CJK Ideograph would have comparable and reasonable representation of IDS.

Presenters:

Track 3 - Exciting International Features of Windows 8

Michael S. Kaplan
Program Manager, Microsoft

Windows 8 has made some exciting investments in the world, its standards, and its languages. This will be one of the very first chances to hear about a lot of them, before anyone else. You won't want to miss this early glance at the newest version of Windows, shipping soon after the conference is done!

12:00-13:00 - LUNCH

SESSION 10

Presenter:

Track 1 - ICANN's Work on Internationalising The Domain Name System

Kim Davies
Manager, Root Zone Services, ICANN

The Internationalized Domain Names for Applications (IDNA) protocol has introduced new possibilities for Unicode strings to be used to expand the available set of Domain Names used on the Internet. ICANN, as the coordinator of the Domain Name System (DNS) root zone and other aspects of the DNS, has overseen the expansion caused by the introduction of internationalized domain names.

.

ICANN's experience has not been without its difficulties. ICANN will explore the history of introducing internationalized domain names to the DNS root zone, and its practical experience deploying the technology.

In particular the ICANN community is wrestling with the complexities where multiple Unicode strings may be considered to be equivalent. The DNS is not well suited to the concept of many potential domain names mapping to one. However, to support the linguistic nuances of many languages and scripts, this concept might need to be supported.

The solution is likely to involve a mix of technological improvements, operational improvements, and reconsideration of the basic premise of "What is a Domain Name?" It is a complex problem, and the discussion of the solution space thus far has not been as widely participated in as it probably should. To find the best solution for the community, ICANN is seeking to broaden participation in discussing solutions in order to come to agreeable approaches. The solutions need to address the gap between what is technically possible, and what is the best way to enable the intuitive use of Internet identifiers by end users.

Another dimension to the roll out of internationalized domain names is striving to ensure once a domain name exists, that it functions in most, if not all, software. This means widespread implementation of the IDNA protocol in applications, as well as breaking implementor's assumptions like "All domain suffixes are two- or three-characters long". Despite the significant growth in the number of Top-Level Domain names (TLDs) in the past five years, and being on the precipice of launching a significantly larger number in the coming years, we still see software vendors taking missteps like hardcoding lists of valid TLDs in their software.
This presentation will review the history of the ICANN community's work to internationalise the Domain Name System, examine the current main issues of TLD variant-handling and universal acceptance that hinder further adoption, and make an appeal for assistance from the Unicode community to help inform and improve ICANN community's work in this area.

Presenter:

Luke Swartz
Product Manager, Google Inc.

Aharon Lanin
Software Engineer, Google, Inc.

Roozbeh Pournader
Internationalization Engineer, Google, Inc.

Track 2 - Bidi and RTL Language Technology at Google

Right-to-Left languages pose interesting internationalization problems in software development. This talk will discuss recent innovations developed at Google to help deal with some of these problems, such as wrapping of Bidi text in Closure Templates and automatic CSS mirroring; many of these innovations have been released to the open source community. We will also introduce some new Bidi challenges that are associated with software developed for mobile devices.

Presenter:

Rakesh Lal
Globalization QE manager, Adobe Systems

Srijan Sandilya
Globalization Test

Track 3 - Testing Challenges for RTL Languages Support in Adobe Products

The Topic will cover the entire spectrum of Challenges and mitigation for extending the support for MENA languages in Adobe Products. The support for Arabic and Hebrew was added by adobe in CS6 for Photoshop, Illustrator, InDesign, Dreamweaver and Acrobat. Prior to this, these languages were taken care of by Winsoft. The challenges among other things included vendor selection, Beta testing strategy, RTL support testing, Font and Keyboard support.

lead, Adobe systems

The presentation will begin with an introduction of the MENA market and then will look into each of the challenges that we faced and how we mitigated them based on extensive market research and pre-release feedback. Adobe introduced specific support for Arabic and Hebrew typography which was tested extensively by both Native and non-native testers with co-ordination from the Adobe team. It is common to use an English OS and install an Arabic Application, the cross-locale user experience had to be thoroughly tested and it posed it own challenges.

The presentation can be useful for companies who are looking to add MENA locale support to their products and also to Vendors who would like to understand the challenges faced by the Product companies.

Here are the section We would like to cover in the presentation:

- Introduction of Presenter
- MENA Market
- What countries group to form MENA region
- Publication industry and financial trends and growth in these markets.
- Challenges in BiDi testing
- Logistical
- Language complexity
- Testers
- Vendor selection
- Beta program
- Technical
- Market feedback
- i18n
- Font support
- Keyboard and IME support

14:00-14:50	SESSION 11

Presenter:

Martin J. Dürst
Professor, Aoyama Gakuin University

Track 1 - Update on Internationalized Resource Identifiers (IRI) and Email Address Internationalization (EAI)

On the Internet, exchanging content (documents, web pages, e-mail messages and the like) in the world's languages and script can be taken for granted. On the other hand, identifying content and people using anything but a very limited set of Latin characters is still difficult if not outright impossible. For URIs (Uniform Resource Identifiers), this is being changed by IRIs (Internationalized Resource Identifiers). For electronic mail addresses, this is being changed by EAI (Email Address Internationalization). IDNs (Internationalized Domain Names) are used in both cases. This paper will give an update on the latest developments in the area of internationalized identifiers. This update will discuss motivations and limitations, basic approaches and methods, the current state of implementation and standardization, and unsolved problems.

For all kinds of identifiers that extend beyond non-accented Latin characters, Unicode is indispensable, but different kind of identifiers take a different approach to using Unicode. For processing in the DNS proper, IDNs use punycode, a compact but cryptic and therefore often despised "ASCII-compatible" encoding. IRIs use UTF-8 and %-encoding when downgrading to URIs. EAI is the most straightforward: In a bold move that would have been unthinkable a few years ago, it extends the underlying e-mail infrastructure to "just use UTF-8".

When this abstract was written, the most serious unsolved problem relating to internationalized identifiers was bidirectional identifiers, i.e. identifiers including characters written right-to-left such as Arabic and Hebrew. We will report on progress in this area, even if we don't expect it to be completely solved by the time of the conference.

This presentation is of interest to anybody working on or using e-mail addresses or web addresses in an international setting. It is based on the hands-on implementation experience of the authors.

Presenter:

Adil Allawi
Director, Diwan Software

.

Track 2 - Socializing Bi-Di

At the 33rd Unicode Conference, I outlined a way to markup social web messages so that urls, # tags and @ mentions can be rendered correctly with right-to-left languages like Arabic. Amazingly, Twitter listened and this method now forms the basis of their current right-to-left support. This paper will talk about how that method works and what is needed and being done to make this a general method for the whole of the web.

Social networks like Twitter and Facebook have had a huge influence on modern langauge. New words and phrases like LOL, ??????, :D, @mentions, #tags and URLs are constantly being introduced and becoming part of common useage. But for bi-directional languages like Arabic this presents a major challenge as these phrases cannot combine easily with right-to-left text.

The Unicode Bi-Di Algorithm has been a great benefit for software in general and forms the basis of right to left text support on the web. It provides a unified way for rendering mixed right-to-left (e.g. Arabic) and left-to-right (e.g. English) text across all kinds of software and devices. However, it cannot cope with the new uses that characters are being put to.

The solution is to use addtional HTML markup (e.g <span dir="rtl">) and/or Unicode directional characters (e.g. U+200F, right-to-left mark) to define the directional properties of segment of text. But what happens when this text is taken from one web site and fed into another or archived and retrieved in the future? The text and numbers can become unreadable, URLs may be unusable or the meaning of a tweet changed. It can be hard to predict how to mark-up the integrated content for the right result. This presentation will cover real world issues and attempt to suggest practical solutions.

I will go on to outline a method to markup bi-directional social text and how this has worked for Twitter. I will continue with suggesting how this can be made into a general solution across the web and how directional information can be preserved as text is transmitted from browsers to servers and back again.

*At the lowest level there needs to be a parser to spot URL's and wrap these correctly. A parser to spot brackets and make sure the open bracket matches the direction of the close bracket.
*The next level up would be a standard way to guess if a stream of text or HTML is primarily right-to-left or left-to-right.
*And the last level is agreed standards for Social API's, XML feeds, XSL transforms that define the intention of the creator of the content.

The presentation will conclude with a proposal for a standard approach and hopefully one that can be supported by all the web sites.

Presenter:

Richard Gillam
Senior internationalization engineer, Lab126

Track 3 - Pseudotranslation: Where the Rubber Hits the Road

The concept of pseudotranslation is simple enough, but actually implementing it in a real live system can be dauntingly complicated. When I joined Lab126 a year ago, we were gearing up for the release of the first internationalized Amazon Kindle, and I was asked to put together a pseudotranslation tool to help our QA team make sure everything was properly initialized. This turned into an adventure. This talk will explore the many challenges and pitfalls we encountered in implementing a new pseudotranslation tool and process from scratch and applying it to an existing system.

Among the challenges we faced:

- We had several different types of resource files in the system
- Many individual resource strings didn't contain user-visible text and didn't need to be translated.
- Our Java ListResourceBundles often used the full generality of Java, making them very complicated to parse and translate correctly.
- Our build system needed to be extended to build and run the pseudotranslation tool.
- Our requirements for the actual content of the pseudotranslated strings shifted over time and needed to be flexible.

14:50 – 15:10 - Afternoon Refreshments

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

Presenters:

Track 1 - Who Are You? User Identity and Unicode

Yoshito Umaoka
Software Engineer, IBM

User authentication and access control in a heterogeneous system requires special consideration for verifying user's identity. A typical system utilizes an LDAP server for checking user's identity and access controlled materials are checked with the resolved identity. The LDAP specification v3 supports the use of Unicode characters. However, the algorithm used for comparing text is not consistent across LDAP server implementations. This may expose a critical security issue. This session discusses the nature of text comparison in directory systems and potential security issues around the technology.

Presenter:

Track 2 - Building the Phonetic Keyboard for Cherokee

Michael S. Kaplan
Program Manager, Microsoft

It has been over a decade since any keyboard layout on Windows has raised the roof on what a keyboard can do, but with heavy use of the long-existing and free but never before used feature of chained dead keys, the single most complicated layout ever created for Windows was created, in a specific answer to a customer request ignored on every other platform. This feature opens the door to many other novel and interesting keyboard layouts that future versions may well find themselves taking advantage of!

Presenter:

Track 3 - I18n Testing for Social -- Case Study from Google+

Katsuhiko Momoi
Staff Test Engineer, Google, Inc.

Social apps are becoming quite popular and prevalent. Applications like Twitter, Mixi, Facebook, Google+ are attracting ever increasing number of users. Social apps can be quite varied depending on how they are architected and what they aim to achieve. Thus needs for i18n testing may cover a wide range of features and dimensions that may not be present in non-social applications. Social apps are also particularly suited for mobile environments and if anything mobile apps are likely to eventually dominate over desktop apps. Drawing on the experience of working on Google+ international testing from the inception of the project in early 2010, I will discuss how we have developed i18n testing strategies for ever expanding and growing list of features and components in a single project. What worked and what did not and how to plan for the i18n testing of a product with rapidly changing UI and features.

First a product overview. Google+ is a single product but it has a number of core features and many associated components. And there are likely to be other additional components in future. In a project like this, there is need for many i18n testing teams or else each component/feature testing team needs to absorb i18n testing. There are some complicating factors. For example, while there are separate component and features, they all integrate into the core application part of Google+. Thus i18n testing needs to address issues arising from integration such as ensuring that each component delivers the same language content as the main app. We will discuss how we approached an overall i18n testing in this environment.

Second. There were a number of features that were not encountered in earlier Google applications. We needed to come up with new i18n test strategies for them. These features include items like Profile name formats for different languages, age limitation policies for different countries, adult status qualification for Google+ Pages, cross-product language settings, mail notification language, user gender and its effects on some features, etc. There was new name checking code put in place to generally classify profile names as inappropriate, violations, etc. There were also human reviews of name violations under language/culture specific conditions.

Third, there were new i18n libraries/APIs that were used for the first time in Google+. The notable ICU libraries were Plurals API and Gender API. These libraries went through first time implementation issues with Google+. There were difficulties encountered by developers and translators who translated strings new message formats. In this presentation I will catalog some of the major issues found during testing and what changes were implemented to address them.

And lastly, we used various automation methods and tools. Some tools like pseudolocalized builds and automatic translation status checker were particularly effective during an early phase of the development. On the other hand, automated tests for UI and layout were largely ineffective at this stage. We encountered several issues with automatic login tests under several languages. In Google+ environments, automated tests are to be run continuously against code under development. Some will be used as a set of presubmit tests. Thus running time is of paramount importance. Under such a test environment requirement, running tests under multiple locales is considered costly. Thus we needed to balance the need for i18n testing against the efficiency of the test system. One major difference between the mobile and desktop testing is that for the former, layout breakage is more often encountered. Yet it is not easy to come up with effective i18n layout testing tools. I will discuss layout testing and possible approaches.

16:10 - 17:00 **SESSION 13**

Presenter:

Kshitij Gupta
Computer Scientist, Adobe Systems

Track 1 - Attacking Globalized Software – and preventing it!

The session is targeted at the audience interested in-depth learning about Unicode encodings and its related security issues. Both intermediate and expert globalization professionals will benefit from attending the session and will be able to identify the real world security loopholes around character encodings. By the end of the session, the audience would have a thorough understanding of the security concerns around character encodings, the best practices for developers & users to prevent such issues, and the industry standards for multiple character rendering algorithms.

Session Agenda and key take-aways:

The session begins with the brief overview of the real world case studies around the hackable web domains owing to the Unicode security loopholes. We demonstrate to the audience how the real world web domains are prone to being hacked with slight modifications to the domain name character set.

A brief history of Unicode and globalization follows with focus on the character encodings and myths surrounding Unicode encoding are discussed with the audience. This is followed by a detailed analysis of the security loopholes and the way to scrutinize software to plug these security loopholes. To better present security concerns, complete list has been classified into visual and non-visual security issues. These classifications are further divided into multiple categories to provide a detailed overview to the audience.

The industry best practice guidelines for security considerations in the globalized applications are presented to participants. Along with the industry standards, we also provide the audience with our recommendations for both the users and the developers. An open discussion to resolve the audience queries follows to close the session.

Note: If time permits, we can include a detailed analysis of Unicode standards like: 'The Bidirectional Algorithm', 'Unicode Normalization Forms' among other standards.

Why this session is important for the Unicode Conference:

The session falls completely in line with the central theme of the conference 'Unicode' and provides the audience a unique opportunity to learn the details of hacks around the globalized software, with the help of some real world case studies and the detailed analysis of the multiple standards for using the Unicode character set.

Presenter:

Raymond Wainman
Software Engineering Intern, Google, Inc.

Cibu Johny
Software Engineer, in Google Inc.

Mark Davis
Sr. Internationalization Architect, Google Inc.

Track 2 - LDML Keyboard Structure Proposal

Keyboard layouts for many languages and scripts have been around for a long time. With the advent of touch sensitive smart phones, tablets and cloud input tools there is an ever growing variety of keyboard layouts available to users. This calls for a standard way to represent the underlying keyboard behavior to be used in a variety of platforms, technologies and visual layouts. Such a standardized format allows for the communication of keyboard mapping data between different modules, and the comparison of data across different vendors and platforms. This presentation will describe the proposed XML specification for keyboard layouts which is currently being reviewed by the CLDR committee. In addition, we have also collected data from the major platforms like Windows, ChromeOS, Mac OSX and Android to help support the specification. Included in the proposal are mappings between keys and outputs, deadkey behavior and long-press options. Advanced IME features, handwriting recognition and the precise physical layout of the keyboard are out of scope of this proposal.

Presenter:

Tex Texin
Chief Globalization Architect, Xencraft

Track 3 - Does it Hurt When I do This? Data for I18n Testing

This presentation recommends specific data values that are likely to identify internationalization problems in software intended for global markets.

Based on years of global software experience, these data values are useful in functional or linguistic QA tests of internationalized software. The data value recommendations include character encoding, postal address, locale and other data types typically used in software and will assist in finding common internationalization problems. This presentation will offer specific test suggestions.

Program is subject to change.

- To Register for IUC36: http://www.unicodeconference.org/registration.htm
  Or, contact Lisa McAdam" at lisa@omg.org
- Exhibitor Information: http://www.unicodeconference.org/be-exhibitor.htm
  Or, contact Ken Berk at ken.berk@omg.org
- Sponsor Information: http://www.unicodeconference.org/be-sponsor.htm
  Or, Ken Berk at ken.berk@omg.org, or 781-444-0404.

Object Management Group®, (OMG®) organizes the Internationalization and Unicode Conferences around the