UNICODE®

Internationalization & Unicode® Conference

40

NOVEMBER 1-3, 2016 • SANTA CLARA, CA USA

CONFERENCE PROGRAM

Tuesday, November 1, 2016

08:30-10:00	SESSION 1 TUTORIALS

Presenter:

Track 1: An Introduction to Writing Systems; Unicode - pt. 1

Richard Ishida
Internationalization Activity Lead, W3C

This tutorial helps you understand the unique characteristics of non-Latin writing systems that impinge on the implementation of Unicode-based applications. It doesn't provide detailed coding advice, but focuses on essential requirements you must understand to deploy Unicode-based solutions, and does so across a representative range of all the world's scripts (including Chinese, Japanese, Korean, Arabic, Hebrew, Thai, Hindi/Tamil, Russian and Greek). It also provides memorable examples to help you understand the buzzwords used in the rest of the conference and your future work with Unicode.

The tutorial starts with basic character encoding principles, but goes much further, covering things such as 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. It has a proven track record as an orientation for newcomers to the conference, but also appeals to people at intermediate and advanced levels, due to the breadth of concepts discussed and the way they are related to real-world script usage. No prior knowledge is needed.

Presenter:

Track 2: Tailoring Collation to Users and Languages

Markus Scherer
Unicode Software Engineer, Google, Inc.

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 – are explained and their effects demonstrated. Discuss will continue with 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 will discuss implementation techniques at a high level, but no source code shall be harmed during this session.

Presenters:

Roozbeh Pournader
Internationalization Engineer, Google, Inc.

Mihai Nita
I18n Sr. Software Engineer, Google Inc.

Track 3: Android Internationalization

A tour of Android’s internationalization and localization features, including a tutorial for developing an internationalized Android app from scratch (localizability, formatting, bidi, etc.). New internationalization-related features of Android N will also be discussed, especially the new support for multilingual users.

10:00-10:30 - Morning Refreshments

10:30-12:00 SESSION 2 TUTORIALS

Presenter:

Richard Ishida
Internationalization Activity Lead, W3C

Track 1: An Introduction to Writing Systems & Unicode - pt. 2

This tutorial helps you understand the unique characteristics of non-Latin writing systems that impinge on the implementation of Unicode-based applications. It doesn't provide detailed coding advice, but focuses on essential requirements you must understand to deploy Unicode-based solutions, and does so across a representative range of all the world's scripts (including Chinese, Japanese, Korean, Arabic, Hebrew, Thai, Hindi/Tamil, Russian and Greek). It also provides memorable examples to help you understand the buzzwords used in the rest of the conference and your future work with Unicode.

The tutorial starts with basic character encoding principles, but goes much further, covering things such as 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.

It has a proven track record as an orientation for newcomers to the conference, but also appeals to people at intermediate and advanced levels, due to the breadth of concepts discussed and the way they are related to real-world script usage. No prior knowledge is needed.

Presenters:

Steven Loomis
Software Engineer, IBM

Yoshito Umaoka
Software Globalization 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. Topics covered will include packaging of ICU data and integrating ICU into an application's development process.

Presenter:

Track 3: Web Internationalization

Tex Texin
Globalization Architect, Xencraft

This tutorial, updated in 2016, is an introduction to internationalization on the World Wide Web. The audience will learn about the standards that enable 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.

12:00-13:00 - LUNCH

13:00-14:30 SESSION 3 TUTORIALS

Presenter:

Track 1 - Internationalization: An Introduction

Addison Phillips
Globalization Architect, Amazon

What is internationalization? What do developers, product managers, or quality engineers need to know about it? How do you incorporate internationalization into the design, implementation, and delivery of a software product? This session provides an introduction to the topics of internationalization, localization and globalization. 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.

All new and revised for IUC 40.

Presenters:

Track 2 - The CLDR Tutorial - pt. 1

Steven Loomis
Software Engineer, IBM

John Emmons
Senior Software Engineer, IBM

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.

Presenters:

Tex Texin
Globalization Architect, Xencraft

Mike McKenna
I18n Product Owner, PayPal, Inc.

Craig Cummings
Principle Software Engineer, Informatica

Track 3 - Introduction to Unicode and Beyond

The abstract that follows is for the 'Introduction to Unicode and Beyond' tutorial with the same presenters as for IUC39. That is, Tex Texin, Mike McKenna, and Craig Cummings.

This tutorial will give you the knowledge for correct implementation for using Unicode to process text in any language. Unicode is the text encoding standard covering every major language on the planet.

Taught by software internationalization experts, this tutorial will introduce you to the key principles of Unicode, its design and architecture, and provide you with examples of real world implementation. Attendees will come away with a basic knowledge of Unicode and how to be more effective at processing, handling, and debugging multilingual text content.

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

14:30-15:00 - Afternoon Refreshments

15:00-16:30 SESSION 4 TUTORIALS

Presenter:

Anshuman Pandey
Language Technologist

Track 1 - 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.

Track 2 The CLDR Tutorial - pt. 2

Presenters:

Steven Loomis
Software Engineer, IBM

John Emmons
Senior Software Engineer, IBM

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.

Track 3 - Unicode in Action

Presenters:

Tex Texin
Globalization Architect, Xencraft

Mike McKenna
I18n Product Owner, PayPal, Inc.

Craig Cummings
Principle Software Engineer, Informatica

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.

Wednesday, November 2, 2016

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

09:15-10:00 **KEYNOTE PRESENTATION - My Life as a Higher Level Protocol**

Presenter:

John Hudson
Co-Founder, Tiro Typeworks

John Hudson has spent two decades working at the messy interface between text encoding and typography, much of it making fonts for complex scripts.

In this keynote presentation, he reflects on some of the messiest aspects of this work, in those places where the character/glyph distinction breaks down, and where it isn't always clear who is responsible for getting text to look the way the author intends and the reader expects. After twenty years, he's convinced that a holistic overview of text is necessary, one that takes in all the steps from encoding to input to visual articulation.

10:00-10:30 - Morning Refreshments

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

Presenter:

Track 1 - How to Ensure Globalization Readiness Through UX/UI Analysis in a Fast-Paced Environment

Matteo Taddei
Independent

User Experience and User Interface Design are a key aspect of 21st century products and are usually positioned very early in the Product Development Life Cycle (PDLC). Very often, especially in an Agile environment, we can notice a strict correlation between User Stories and UX/UI mockups. In this talk we will cover how to ensure your product is Global 1st starting from the Design Conception phases through static analysis of Design mockups. We will define clear rules for static analysis, explicit recommendation based on CLDR and Unicode to limit the number of country-specific customization keeping in mind the importance of being both Global and Mobile 1st.

Presenters:

Track 2 - New in ICU

Markus Scherer
Unicode Software Engineer, Google, Inc.

Steven Loomis
Software Engineer, IBM

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.

This presentation will provide an overview of ICU, with emphasis on the recent updates in ICU 57 & 58, including the latest support for Unicode 9.0 and CLDR 29/30, line breaking improvements, new date/time formatting capabilities, and other changes. The presentation will also touch on ICU’s planned direction for future releases.

Presenter:

Shivangi Rai
Technical Staff, Adobe

Track 3 - Modern Web App Internationalization: Techniques & Frameworks

English is not the first language for over 70% of Internet users. For a Web App to be successful it must be well localized and internationalized. If your business is (or could be) global, this session will give you more insight and skills to create a well internationalized modern web application built with AngularJS, ReactJS and EmberJS.

This session is ideal for those developing a modern web application based on libraries like AngularJS, ReactJS and EmberJS. Majority of businesses have moved onto the web. Companies are investing a lot in modern web apps. Since, a significantly large part of the market is non-English speaking, so developing a product which is well localized and provides multilingual support has become important for growth of such products and in turn the related global businesses. Modern Web Application Development is many based on MVC architecture. AngularJS, RecatJS, EmberJS are JavaScript frameworks that implement MVC.

You normally perform following steps to make an internationalized product:

1. UI strings externalization
2. Unicode Support
3. SEO support
4. Unicode Text Editing Tool - provides support for complex scripts.

In this session, we will see the learnings from internationalization of Adobe Story Web which is a collaborative script development tool. It is developed using angular JS framework. We will see how we are using Etherpad, Angular Translate for internationalizing Adobe Story.

11:30-12:20

SESSION 2

Presenter:

Elsebeth Flarup
Senior Engineering Project Manager, Guidewire Software

Track 1 - Using Anti-Patterns in Internationalization Training

Through more than 20 years part of my job has been to educate software developers and testers on the importance of doing internationalization right, how to do it, and especially what NOT to do. I have done this using the usual code examples, and showing/explaining why this doesn't work in other languages or locales (because they use a different date format, or need a different word order, etc.). In many cases this does not necessarily hit the target audience with a full understanding of how big of an impact bad internationalization can have, however. So I have switched to using anti-patterns in my training: creating mocked up UI screens that show how incorrect internationalization would look in English, in an en-US locale. So far experience my experience has shown that this creates a much higher degree of engagement during training. In this session I will demonstrate an example of this, and how I tie the initial anti-patterns to examples pulled from actual code.

Presenters:

Vivek Pani
Co-Founder, CTO, Reverie Language Technologies

Track 2 - Cross-Language Computing for Indic Languages

Why India needs cross-language computing
Most Indians are polyglots and are comfortable using more than one local language at a time. However, only 10% of them are conversant in the English language. With millions of local-languages users willing to be a part of the Indian Internet, developing cross-language computing tools makes leeway for local-language digital content

discovery and consumption.

Challenges in cross-language computing for Indic languages
At a fundamental level, the existing input methodology followed for Indic scripts allow room for ambiguity and undesired characters. This in itself complicates the cross-language information discovery and retrieval.

Further, the lack of basic NLP tools like stemmers, lemmatizers, spelling correctors, and PoS taggers dilutes the ability to index and retrieve content in multiple languages. Language detection across multiple scripts and transliteration schemes are an impending challenge.

Moreover, less than 0.1% of digital content on the Internet is in languages other than English, which makes it extremely difficult for researchers to build machine learning models.

In this session you will learn:

- Approaches for correcting spelling and transliteration errors in a multilingual environment
- Building NLP tools for Indic languages: Challenges in PoS tagging and Named Entity Recognition
- Building contexts and addressing semantics: Synonomy, Polysemy, and Word Sense Disambiguation
- Towards multilingual content discovery: New approaches to multilingual indexing and relevance ranking
- Language independent NLP: Character Encoding and Deep Neural Networks
- Machine Translation at scale: can we increase content availability and accessibility?

Presenters:

Steven Loomis
Software Engineer, IBM

Yoshito Umaoka
Software Globalization Engineer, IBM

Track 3 - Translating at Cloud Speeds with the Globalization Pipeline

Modern software has development cycles measured in weeks, if not days. Application translation processes have had to constantly adapt in order to keep up with this rapid pace. While agile translation methodologies bring the right mindset, development teams are still cumbered with burdensome manual steps that waste time -- and precious translation dollars.

The Globalization Pipeline was introduced to provide a consistent, easy-to-use hosted service that integrates with your existing development process. It provides access to machine translation as well as human post-editing. Translations are managed within the service without the need to separately manage resource files. Translation updates to source or target languages can be processed through the pipeline either as part of build or packaging steps, or even accessed at real-time for instant translation updates. SDKs to provide seamless integration are available in over seven programming languages and environments, or well-documented RESTful APIs can be accessed directly in additional environments.

This presentation will introduce participants to the Globalization Pipeline and show how it enables modern translated application development. Time will be taken for Q&A on lessons learned and challenges in bringing this service to the public.

12:30-13:30 - LUNCH

13:30-14:20 SESSION 3

Presenters:

Track 1 - Globalization Best Practices at Netflix

Shawn Xu
Internationalization Engineer, Netflix Inc.

Lee Collins
Senior Internationalization Architect, Netflix, Inc.

Representatives from various teams at Netflix discuss the globalization best practices that they follow on day to day basis, developing and supporting a global service enjoyed by over 75 million members in over 190 countries.

Following Netflix culture of Freedom and Responsibility, teams at Netflix have lots of freedom to implement things the way they see fit, but at the same time, it is each team's responsibility to find and follow the best practices. How do we evangelize the globalization best practices and work together with teams to achieve that goal? Come join the representatives from some of these teams to discuss topics ranging from Language Selection Service (a service helps negotiating language for various UIs), to fonts specially designed for Netflix to be used on Smart TVs and devices, as well as unique challenges in adding support for CJK and Arabic.

Presenters:

Track 2 - CLDR New Advanced Topics

Steven Loomis
Software Engineer, IBM

John Emmons
Senior Software Engineer, IBM

The Unicode CLDR has many new types of functionality added in every release. Join us for three informative mini-sessions as we take an in depth look at some of the newer features of CLDR, and how they can be used to provide more complete globalization of computer applications.

Topics to be covered include:

- Validity Data
- Day Periods
- Unit Context and Usage

Presenter:

Track 3 - Real-Time Localization Updates with Web Services

Tarrance Egbert
Globalization Engineer, Adobe

No longer do we need to wait for the the next product build to update translations and other localization information. Using a REST interface, translated strings can be updated immediately for your product.   With the growing number of Saas implementations, it is more and more important to have the ability to provide real-time updates to your content. This session will show how this can be done.

We use a backend translation service to keep strings up-to-date and provide a web service to that backend that allows the website real-time access to new translations. This service can provide a better experience for end-users and software engineers as well. We will demonstrate how these systems work together to provide this service.

14:30-15:20 SESSION 4

Presenter:

Track 1 - Time Zones for International Communications

Addison Phillips
Globalization Architect,

If the time difference between two locations is always a multiple of an hour, and if there is no Daylight Saving

Amazon

Time (also known as Summer Time), then the world can be covered by about 24 time zones. However, the reality is that the time difference can be 30 minutes, or even 15 minutes, and many time zones in the world use Daylight Saving Time. Also, start/end dates of Daylight Saving Time occasionally change. As a result, there are hundreds of time zones in the world. With this level of complexity, it is easy to make mistakes. If someone's mobile device shows a wrong time, he/she may be late for a meeting. This can happen if time zone data is not up to date or if a neighbor time zone with different Daylight Saving Time rule is used instead of the correct time zone. Similarly, if someone posted information about an event on social media and another person could not figure out the time of the event, he/she may fail to join. In the current world of computers and mobile devices connected via worldwide network, understanding of time zones is essential for software designers and developers. This paper explains what the time zones are, how they are defined, and how to properly use them with various examples, so that software designers and developers can show the accurate time to their customers all over the world.

Time zones are defined based on Coordinated Universal Time (UTC) which is the successor of Greenwich Mean Time (GMT). In November 2015, a decision was made to continue using UTC at least until 2023. This paper also discusses brief history/background of GMT and UTC to understand time zones in depth.

Moderator:
Steven Loomis
Software Engineer, IBM

Panelists:
Mark Davis
Chief Internationalization Architect, Google, Inc.

Zibi Braniecki
Senior Software Engineer, Localization Drivers Team, Mozilla

Kristi Lee
Program Manager, Microsoft

Track 2 - 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 a brief introduction to CLDR and what's new in versions 29, and then continue with a panel discussion focused on the experience of direct consumers of (and contributors to) 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 allow plenty of time for questions from the floor, and general Q&A about CLDR.

Presenter:

Debbie Anderson
Researcher, UC Berkeley

Track 3 - Script Encoding - pt. 2: Working with the User Community

Encoding scripts in Unicode involves several steps. Initially, a script proposal must be written that identifies the repertoire of characters with names and glyphs, describes how the characters behave, and gives background and technical details. The Unicode Technical Committee then reviews the proposal – usually multiple times – to ensure adequate information has been provided to be able to implement the script on computers and mobile devices. But the path toward successful script encoding extends beyond the technical aspects of the proposal itself. Proposals need to have the buy-in of the user community, which may comprise scholars and/modern-day language users (who may not necessarily have the support of the government). The users themselves may be split amongst various groups, which adds to the difficulty in arriving at consensus. Problems in communicating and working with the user community can lead to long delays in getting scripts approved. Old Hungarian, for example, took seventeen years to finally get approved and published in the standard. Still, developing solid contacts in the community is vital, and can aid in implementation and font development, as well as in the collection of locale data after the script is published. This talk will discuss the steps involved in working with the user community, citing real-life examples where conflicting views on script proposals arose, and what lessons can be learned for future script encoding. The speaker, who runs the Script Encoding Initiative project at UC Berkeley, will draw from her on-the-ground experience. Questions to be discussed:

- What are the constituencies using the script, where do they live, and what agendas do they have?
- Are there reliable contacts in the user community with email capability, and is language an issue in communicating with users?
- How can one reliably screen contacts to verify they represent the key people in the user community?
- How can one most effectively engage with the user community?
- What to do if no user community can be found or contacted?

15:20-15:50 - Afternoon Refreshments

15:50-16:40 **SESSION 5**

Presenter:

Alolita Sharma
Senior Manager, Internationalization Architecture and Engineering, PayPal, Inc.

Track 1 - The Rise of Emoji

Emoji is taking over the Web.

Emoji is helping users express their feelings -- on social media, election campaigns, restaurant reviews, advertising and much more. We will explore some of the creative ways people are communicating with Emoji. For example, a profound effect of using Emoji is to enable diverse people to connect across languages and cultures.

Understanding Emoji can provide web platforms with deeper insight into their users. We will look at how user generated content platforms by Facebook, Twitter, Yelp, Microsoft, Google are leveraging machine learning, language processing technologies and user data at scale to understand Emoji and related user sentiment.

Presenters:

Dominik Röttsches
Senior Software Engineer, Google, Inc.

Behdad Esfahbod
Internationalization Software Engineer,

Track 2 - Unicode on the Web: Text Rendering in the Chrome Browser

In this session we will talk about the challenges of rendering complex Unicode text on the web. In particular, we will discuss architectural changes that we have made in the Chrome browser over the last two years that resulted in drastic improvements in the following areas:

- Better font selection and fallback, including shaper-driven font fallback and segmentation
- Improved text shaping performance, which allows us to enable
- OpenType processing for all languages (including those based on Latin script) and remove the so-called simple-path shaper

Google, Inc.

- Full support for emoji sequences, better emoji font selection & fallback, and correct emoji line-breaking
- Web typography improvements: Implementing CSS font-variant-* subproperties for easier access to typographic features
- Improved cross-platform font rendering fidelity, while dealing with unique challenges of each platform’s native font APIs
- Improved implementation of Unicode variation sequences

Presenters:

Shawn Xu
Internationalization Engineer, Netflix Inc.

Shashi Mathada
Senior Software Architect, Netflix, Inc.

Shervin Afshar
Localization Engineer, Netflix, Inc.

Track 3 - Localizing Content with Grammatical Plural Numbers

Localizing content for the languages of the world can be fun and challenging, especially so when the topics of grammatical plural numbers get in the picture. At first sight, the documented plural rules for languages of the world can seem mind-boggling. Developers tend to write English source strings as simple as possible and translators are mostly not familiar with complex string structures provided as solution to deal with grammatical plural numbers. Shawn, Shervin, and Shashi will discuss how these challenges were tackled at Netflix. They will cover how a collaboration between multiple teams made it easier for developers to create better English source strings and for the translators to translate with grammatical plural numbers in mind. This talk will walk you through the journey; the training for developers and translators will be covered, tools and technologies built and adapted for the purpose would be presented, and finally lessons learned would be discussed.

16:50-17:40 SESSION 6

Presenter:

Mark Davis
Chief Internationalization Architect, Google, Inc.

Track 1 - Emoji Characters

- What are emoji?
- Where did they come from?
- How are new ones added?
- How do they work?

What‘s happening next?

Presenters:

Martin Dürst
Professor, Aoyama Gakuin University

Alolita Sharma
Senior Manager, Internationalization Architecture and

Track 2 - Lightning Talks

This is a re-installment of the very successful first Lightning Talks at last year's IUC 39. 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 of topic areas listed in the CFP. This is the chance for you as a conference attendee to present your latest idea or development, spread the word, or raise awareness about something of importance to you, or talk about a topic that doesn't need a full session.

Please send proposals for lightning talks to the moderators, Alolita Sharma and/or Martin Dürst, by October 23. If we have any remaining slots, we will also accept proposals during the conference. Questions on any of the lightning talks will be at the end of the 60 minute session.

Engineering, PayPal, Inc.

Presenters:

Debbie Anderson
Technical Director, Unicode Consortium,

Lisa Moore
Technical VP & IUC Conference Chair, Unicode Consortium

Craig Cummings
UTC Vice-Chair, Unicode Consortium

Track 3 - Behind the Curtain: The Unicode Consortium in 2015-2016

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

18:00-19:00 - CONFERENCE RECEPTION

Thursday, November 3, 2016

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

Presenter:

Jim DeLaHunt
Principal, Jim DeLaHunt & Associates

Track 1 - Discourse and OpenToonz: Free Software I18n Case Studies

It's one thing to say how internationalization ought to happen. It's quite another to observe how it actually does happen. These two case studies of recently-opened free software projects let us see how volunteer efforts at internationalization and localization actually emerge.

Discourse is a clean-sheet design for website-hosted discussion forums. It allows seamless access from mobile devices, desktop web browsers, and email. It is built with Ruby and PostgreSQL, which assume Unicode text and i18n support. It is deployed with modern packaging like Docker. The core development team implemented an English version, but volunteers soon provided localizations in several languages.

Toonz is animation software originally from Italian developer Digital Video S.p.A., and extensively modified over the years by Japanese animator Studio Ghibli as an in-house tool, with Japanese language UI and documentation. In March 2016 a Japanese publisher, Dwango, published as the Ghibli version of Toonz, with a free license, as OpenToonz. There was immediate world-wide interest in the project, and a localization into English was a first order of business. In contrast to most software studied at IUC, the OpenToonz project is not fully and primarily in

English; in some aspects, the Japanese language is primary, and English is a localization target.

For both these projects, we look at the extent of localization and internationalization by the original developer. Then we see how demand for localization manifested from the free software community. How did language teams emerge? What false starts were there? How well was the community able to improve internationalization from the grass roots? How well did the baseline internationalization of the development environments and tool yield a better internationalized project before any further efforts by the original developers?

Session outline:

- Overview of Discourse i18n of its component tools (Ruby, PostgreSQL, Docker)
- I18n and l10n by the original development team
- Community-driven l10n, and I18n
- Overview of OpenToonz
- I18n of its component tools (Ruby, PostgreSQL, Docker)
- I18n and l10n by the original development team
- Community-driven l10n, and I18n
- Special observations from it being Japanese-original with English as a localization
- Lessons learned about free-software l10n and I18n from both projects

Presenter:

Track 2 - Complex Text Layout on Simple Devices

Pedro Navarro
Senior Software Engineer, Netflix, Inc.

As part of Netflix's global expansion our UI framework had to support all the world's major scripts, but we couldn't use any of the readily available libraries because of hardware constrains: our code had to be 100% portable among a great variety of devices with different operating systems (game consoles, Android TV, OS X, Linux) and capabilities (very low system and graphics memory or lack of a writable file system). We'll talk about our journey from a simple DirectFB application using FreeType to a custom font and text layout engine that uses Harfbuzz and ICU and the tradeoffs we had to make between features and code size. Mistakes were made, code had to be refactored and deadlines were not met, but in the end we shipped a highly optimized and portable text engine with font fallbacks, markup support, bidirectional and vertical text, emojis (of course!) and some special features needed by subtitles. We'll present our solutions to some of the problems commonly faced by text engines and the challenges we faced because of the lack of standards or clear guidelines for text layout.

Presenter:

Track 3 - How We Built a Keyboard for Polyglots

Daniela Semeco
President, Polyglotte Inc.

The PolyKeyboard® is made for cross-cultural communication and lets you type correctly in multiple languages without slowing down! Our novel keyboard design is a software solution that works with a physical computer keyboard as well as mobile devices. We overcame the challenge of creating a solution that spans all platforms, whether you write on an iPad or touch type.

The idea was born in February 2011. We had a limited budget, and used resources at hand:

- My father built our first prototype, a notepad application for Windows.
- Next, we created a private Web-based prototype.
- Then, our Free iPad demo app launched in October 2013.

In September, 2014 Apple opened its doors to third-party keyboards, and our universal app (for iPhone, iPad, and iPod Touch) hit the market March 30th, 2015. With this new product, we introduced keyboard localization for the first time (QWERTY, AZERTY, QWERTZ).

Our physical keyboard for Windows was upgraded in January 2016.

Our universal app for iPhone, iPad and iPod Touch was improved July 2016 (adding five more languages, plus more symbols for law, math, and science). Our keyboard now supports 25 languages and more than 250 symbols. Type in multiple languages without losing your train of thought, stopping to look up a special character, or Googling a word to copy and paste it into your document. Users are able to choose which of the 25 languages they prefer to write in, and deactivate the others. We have also created a simple one-step learning process that will allow users to quickly take advantage of the new capabilities introduced by our PolyKeyboard®.

10:00-10:50 SESSION 8

Presenter:

Santhoth Thottingal
Senior Software Engineer, Wikimedia Foundation

Track 1 - One Hundred Thousand Translations of Articles: Wikipedia Translation

Wikipedia has a new article translation system with close integration to its editing workflow. Since it was launched more than a year back, it helped creating hundred thousand plus articles to more than a hundred languages. At this point a new wikipedia article is created using this translation tool in every five minutes.

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 suggests articles to translate based on the interests of translators.

This presentation is about the interesting challenges in building such a large multilingual system and how we solved it.

Presenter:

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

Track 2 - Developing a Pan-CJK IVD Collection

The latest version of the IVD (Ideographic Variation Database) includes only registered collections that are specific to a particular region or language. However, variation across regions is a fundamental characteristic of CJK Unified Ideographs, which serves as the premise for developing Pan-CJK fonts. One of my longer term goals for the open source Source Han Sans / Noto Sans CJK project has been to represent regional variation in "plain text" through the use of registered IVSes (Ideographic Variation Sequences), and the main focus of this presentation will be on the proposed "PanCJKV" IVD collection. One of the barriers toward this goal has been the degree to which glyphs can be shared across regions, which largely depends on the typeface style and typeface design. An approach that depends of typeface style or design could easily lead to several incompatible Pan-CJK IVD collections. One solution is to register a general-purpose and future-proof IVD collection that includes a large number of registered IVSes that covers all CJK Unified Ideographs (there are 80,388 in Unicode Version 8.0) and all regions that use CJK Unified Ideographs. Toward this end, I have proposed the registration of the "PanCJKV" IVD collection. The pros and cons of this IVD collection, along with alternate solutions that can be considered, will be explored as part of this presentation.

Presenters:

Track 3 - How to Handle International Keyboard for Virtual Desktops in Cloud

Zhenjun Zhuo
Globalization G11n Consultant, VMware

Qiang Wan
Internationalization Engineer, VMware

Peter Jonasson
Sr. Quality Engineering Manager, VMware

Jim Peng
Senior Manager of Engineering, VMware

Nowadays, users can easily deploy desktops in private, hybrid or public Cloud environments. A single virtual desktop could reside and be utilized anywhere in the world and receive input from a myriad of different international keyboard layouts and different platforms (Windows, Mac, iOS or Android). It is very important to provide a seamless user experience for these scenarios. During this session, we will introduce the keyboard handling mechanisms across different platforms and summarize crucial international keyboard issues plus solutions observed and validated in a cloud environment. Some of the issues presented would be mapping-, browser-, and external keyboard related. In closing we will outline challenges and future works under consideration.

10:50-11:10 - Morning Refreshments

11:10-12:00	SESSION 9

Presenter:

Track 1 - Math Accessibility

Murray Sargent III
Software Engineer, Microsoft

This talk discusses aspects of making Unicode mathematical zones accessible to blind people. Presumably equations that are typographically simple are accessible with arrow keys and with each variable and two-dimensional construct being spoken or felt when the insertion point is moved to it. At any particular insertion point, the user can edit the equation using the regular input methods. But it can be hard to visualize a more typographically complex equation, let alone edit it. Instead, the user needs to be able to navigate a complex equation using a mathematical tree of the equation. More than one kind of tree is possible and this talk compares a tree that corresponds to the traditional math layout used in documents to a tree that corresponds to the mathematical semantics. The former is seen to be preferable for accessibility.

Presenter:

Track 2 - Stable Web Typography Without Fonts

Thomas Milo
Partner, DecoType -

This is a report about a project initiated by the Sultanate of Oman. The project aims to display a searchable and quotable Arabic Qur’ān text on the web in a typographically stable and orthographically flawless form, regardless

Designers of Computer Typography

the operating system, browser or the type of web device. The webQuran project also features interactive Letter Group Shaping and automatic text Shaping. For this purpose Designers of Computer Typography, also known as DecoType, pioneered new computer typography, without fonts.

Presenter:

Track 3 - Node.js Intl All the Things

Steven Loomis
Software Engineer, IBM

Node.js has become a popular platform, using JavaScript on the server, or in other environments outside of its traditional role in web browsers. This presentation will discuss challenges, lessons learned, and the latest status in enabling and making use of the Intl (EcmaScript-402) module support in Node.js, current status and what's next for JavaScript and Node.js globalization, and discuss techniques and best practices for Unicode and international support in Node.js applications.

12:00-13:00 - LUNCH

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

Presenter:

Track 1 - The New PayPal Mobile App

Lucas Welti
Globalization Architect, PayPal, Inc.

These days most of commerce is done online with mobile becoming the key player.

The previous PayPal Mobile frameworks did not support some of the unique PayPal requirements such as providing English for all Markets and supporting the locale language in some specific countries.

It was important for us to ensure that customers had a consistent experience across all devices from web to smart phone, and for that reason a lot of the Dates, Name, Address, Currency Metadata that is shared between our web applications and Mobile Apps, have only ONE source of truth.

This talk will explain how PayPal was able to launch their new Mobile App for iOS and Android and make it available for 145 markets. This was done using new platforms and processes to make it easier to localize and customize. From Design to Development, the Globalization Team was involved from Day 1, providing Internationalization support, Globalization Q&A, reviewing Content and delivering translations.

Presenter:

Track 2 - Validating Estimates for Software

Tex Texin
Globalization Architect, Xencraft

This presentation will help senior managers validate development estimates for globalizing software and offer guidelines to either reduce the effort or prioritize implementation so that it can be phased into releases that maximize market value. This paper is focused on software internationalization, not localization processes.

Developers generally make estimates in good faith. However, estimates are based on innumerable assumptions that should be reviewed. Software development is complex and so there can be more than one way to approach solutions. The initial estimate is often made with some urgency and with presumed optimal or ideal solutions. This presentation will give program managers, engineering leaders, and product managers insights into questions they can ask to validate software estimates and options that can enable more efficient delivery of international software.

Presenter:

Track 3 - Further Advances in Parabix Technology for High-Performance Unicode

Rob Cameron
Professor, Simon Fraser University

Parabix technology as incorporated in icgrep 1.0 achieves Gigabyte per second performance in the context of the full Unicode level 1 regular expression requirements of Unicode Technical Standard #18. We discuss three directions of progress in further application of Parabix technology to Unicode processing requirements. The first is the technical advances in Parabix regular expression technology to achieve high-performance implementation of Unicode level 2 requirements. The second is the development new pattern concepts including Unicode property contexts and reflective matching based on Unicode name similarity. The third is the development of additional dynamic compilation technology to enable further Unicode applications including transcoding, segmentation and canonicalization.

13:50-14:40	SESSION 11

Presenter:

TBA

Presenter:

Track 2 - BiDi in the Wild: Challenges of the Unicode BiDi Algorithm

Moriel Schottlender
Software Engineer, Wikimedia Foundation

The Unicode Bi-Directional Algorithm (Unicode BiDi) is responsible for making writing and typing right-to-left and left-to-right scripts much easier. There is no doubt that its existence has made the lives of right-to-left users much easier, and opened the door for creating bilingual internationalized Web pages. But, depending on the wider context of the string, the rules which govern Unicode BiDi can prove to be insufficient and inaccurate in some less common cases.

Sometimes, the “correct” behavior of Unicode BiDi algorithm is incorrect for the given usage. Sometimes, they are simply insufficient, and sometimes they are outright confusing. Always, right-to-left users deal with those challenges on a daily basis -- sometimes having to change their typing habits to “cheat” the system and get their desired results.

This lecture will demonstrate the more outrageous challenges that happen when using Unicode BiDi “in the wild”, online, and especially in the complex internationalized ecosystem of Wikipedia. What happens when Unicode BiDi is used in the wild? Is it sufficient? When is it failing us, and what can we do to improve it -- or, generally, the lives of our users?

Presenter:

Ivan Provalov
Software Engineer, Netflix Inc.

Track 3 - Autocomplete Multi-Language Search Using Solr

Autocomplete presents some challenges for search in that user's search intent must be matched from incomplete token queries. Many non-Latin character based languages have additional complications. The following are some of the examples of unique language-specific issues which must be addressed in search systems in order to support these languages:

- Japanese and Chinese multiple scripts (Hiragana, Katakana, Romaji, Zhuyin, Paoding)
- No token-delimiters for Japanese and Chinese
- Korean character composition
- Arabic spelling variations of the transliterated foreign words

I will talk about these challenges in detail, describe our approaches to solving them, and share some tools (queries testing framework) we used to help addressing these issues.

14:50 – 15:10 - Afternoon Refreshments

15:10 - 16:00	SESSION 12

Presenter:

Addison Phillips
Globalization Architect, Amazon

Track 1 - Web Standards: What's Happening?

Standards activity at the W3C and other standards bodies is helping improve the globalization of the Web. There have been impressive improvements in the last few years in support for important languages (as well as ongoing additions for minority ones), new features that bring the Web closer and closer to the richness of print, and developments in areas like IoT (the "internet of things") that affect internationalization professionals and users wherever they use the Web.

This presentation summarizes developments, particularly the activity of the W3C Internationalization Working Group, exploring developments over the past year, as well as the status and challenges of on-going work.

Presenter:

Su Liu
Globalization Architect, IBM

Track 2 - Concept of Cloud Based Globalization Testing Verification Test Service

This paper discusses a cloud based globalization testing service concept and a related framework architecture (prototype). The framework includes a client side globalization feature abstracter (CSGFA) and a server side globalization test application programming interface (SSGT-API). CSGFA collects unstructured globalization related information data (such as, enabled languages, locale names, routines, APIs, and message buffer sizes etc…) from a client application development environment. And, CSGFA converts the collected globalization related information data to a suitable data form, and upload it to SSGT-API as a globalization testing service request. In response to receiving suitable data form, SSGT-API provides globalization verification and examination as a testing service, and then return a globalization verification analysis report. The paper further discussed concepts of prototypes for examining logic relationship between a set of product development data and correlated a set of globalization parameters and globalization dependent APIs. By processing the set of globalization data using a set of globalization testing operations, a globalization test output can be determined and predicted. In response to the determining, the globalization test output can be provided.

Presenter:

Track 3 - Unicode, CLDR, and KYC, KYB, and Money Laundering

Mike McKenna
I18n Product Owner, PayPal, Inc.

In today's world, when money is passed from one party to another, someone else is watching that transaction and must decide if it is legal or suspicious. Every legal money institution is required to "Know Your Customer" - known in the industry as "KYC", and "Know Your Business" (KYB). This is to reduce risk, reduce money laundering, and for regulatory requirements, to comply with local and international governments.

In this session, you will get a quick overview of KYC, KYB, Risk and Compliance, then we will delve into the internationalization aspects and how Unicode and CLDR help. Topics to be shared:

- What is KYC, KYB?
- Risk and Compliance - the what and why
- Anti-Money Laundering
- What are Designated Individuals
- User information and onboarding to applications
- What is unique for KYC and KYB from country to country
- What's in a name? Legal names and login IDs around the world
- How to use CLDR to reduce Risk in user information
- How Unicode helps to reduce Risk in user sign-up
- How International Components for Unicode (ICU) helps with AML and finding Designated Individuals

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

Presenter:

Track 1 - Internationalization in Ruby 2.4

Martin Dürst
Professor, Aoyama Gakuin University

Ruby is a purely object-oriented scripting language which is easy to learn for beginners and highly appreciated by experts for its productivity and depth. This presentation discusses the progress of adding internationalization functionality to Ruby for the version 2.4 release expected towards the end of 2016. One focus of the talk will be the currently ongoing implementation of locale-aware case conversion.

Since Ruby 1.9, Ruby has a pervasive if somewhat unique framework for character encoding, allowing different applications to choose different internationalization models. In practice, Ruby is most often and most conveniently used with UTF-8.

Support for internationalization facilities beyond character encoding has been available via various external libraries. As a result, applications may use conflicting and confusing ways to invoke internationalization functionality. To use case conversion as an example, up to version 2.3, Ruby comes with built-in methods for upcasing and downcasing strings, but these only work on ASCII. Our implementation extends this to the whole Unicode range for version 2.4, and efficiently reuses data already available for case-sensitive matching in regular expressions.

We study the interface of internationalization functions/methods in a wide range of programming languages and Ruby libraries. Based on this study, we propose to extend the current built-in Ruby methods, e.g. for case conversion, with additional parameters to allow language-dependent, purpose-based, and explicitly specified functionality, in a true Ruby way. Both the design as well as the implementation of the new functionality for Ruby 2.4 will be described.

This presentation is intended for users and potential users of the programming language Ruby, and people interested in internationalization of programming languages and libraries in general.

Presenter:

Katsuhiko Momoi
Staff Test Engineer, Google, Inc.

Track 2 - Release Criteria and Mobile I18n Testing Tools

In IUC38 and IUC39, I reported on our efforts to create and promote mobile i18n testing tools. These efforts are continuing into 2016 and we are adding more tools. In this IUC40 talk, I will report on a new project to define i18n release criteria for mobile projects.

There are potentially many items for i18n and localizability testing and defining release criteria referring to all of them is not a simple task. Thus the first step we took is to define the basic i18n release criteria that our products should meet before releasing a product to the public. In setting up the criteria, we have taken a number of factors into consideration. The criteria must be fairly simple to follow; they should be relatively small in number (fewer than 15 to 20); there should be tools or easy to follow instructions to verify a project has met each criterion; make most if not all criteria quantifiable by statistical measures and be presented visually also; and others.

In this talk, I will discuss the criteria that we have proposed and how they are verifiable by the tools we have created and will be creating. Many of these tools have been already open-sourced or will be in the near future. I will also discuss what issues have arisen while trying to use the tools in this task. One important aspect of getting release criteria adopted by product teams is that getting i18n consideration into the design and process of a product development is critical for successful deployment. Thus I argue that getting projects to adopt good i18n engineering practices is the key to deploying these criteria -- in that fulfilling them would be a natural consequence of good i18n engineering practices on a daily or regular basis.

Presenters:

Ballav Bihani
Senior Software Architect, Netflix Inc.

Jose Moreno
Senior Software Architect, Netflix, Inc.

Prosenjit Bhattacharyya
Engineering Lead, Netflix,

Track 3 - What Language to Show?

With a global user base, serving up a localized experience is imperative. It is extremely important that user interactions happen in the right language. For Netflix, be it a new member signup or an existing member searching for titles, or reading the title synopsis on the site or on her device, language plays a very important role. The challenge is to select the correct language based on a user's account preferences, device settings, geolocation, and device capabilities. In this talk, we will discuss how we solve these challenges in Netflix. A few types of problems that we face with when determining the most appropriate language are:

- A user is trying to sign up from Hong Kong, but their device does not support Traditional Chinese.
- What if we don't have assets to show in a particular language?

- What language should we fall back on without causing user dissatisfaction?