CONFERENCE PROGRAM

Wednesday, October 16, 2019

09:00-10:30	SESSION 1 TUTORIALS
Presenter:	**Track 1: An Introduction to Writing Systems & Unicode Part 1**
Richard Ishida, *Internationalization 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 concepts and 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:

Track 2: Introduction to Unicode and Beyond

Craig Cummings
Staff Consultant/ Evangelist, VMware

Mike McKenna
Globalization Strategist, PayPal, Inc.

Tex Texin
Chief Globalization Architect, Xencraft

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

Presenters:

Track 3: Put ICU to Work

Steven Loomis
Senior Software Engineer, IBM

Shane Carr
Senior Software Engineer, Google, Inc.

This tutorial gives attendees everything they need to know to get started with working with Unicode text in computer systems using the International Components for Unicode library (ICU). ICU is a very popular internationalization solution, and is hosted by Unicode itself. While it vastly simplifies the internationalization of products, there can be a learning curve.

The goal of this tutorial is to help new users of ICU install and use the library. The tutorial will walk through code snippets and examples to illustrate 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, or 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 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 with the

fluent API, calendars, collation, break iteration, Unicode properties, transliteration). We will also cover the packaging of ICU data, integrating ICU into an applications development process, and how to get involved in the ICU development community.

10:30-11:00 - Morning Refreshments

11:00-12:30	SESSION 2 TUTORIALS
Presenter: **Richard Ishida,** *Internationalization Lead, W3C*	**Track 1: An Introduction to Writing Systems & Unicode Part 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 concepts and 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: **Craig Cummings** *Staff Consultant/ Evangelist, VMware* **Mike McKenna** *Globalization Strategist, PayPal, Inc.* **Tex Texin** *Chief Globalization Architect, Xencraft*	**Track 2: Unicode in Action** The Unicode in Action tutorial is a 90 minute session that demonstrates programming with Unicode and related best practices. This tutorial will build a simple application and demonstrate the code and resulting behavior as internationalization functions are added. Attendees will be able to relate these prototype examples to the requirements of their own applications and reference them to code solutions. The program will show sorting of different strengths, regular expressions, Unicode normalization, bidirectional languages, and other features of the Unicode standard. The tutorial will highlight why each of these functions are needed so you can determine when to use them in your applications.

Presenter:

Ben Yang
Director of Technology, PanLex

Track 3: ICU Does That? Lesser-known Features of ICU, How to Use Them, How to Extend Them, and How it All Relates to the CLDR

The Unicode CLDR and ICU are widely used for a large number of internationalization and localization tasks. However, there are a significant number of components within them that are not as well known or used. This tutorial will demonstrate the principles behind two of these components:

- Script Transliterators, which allow users to convert a string from one writing system to another (for example, converting the Japanese string "キャンパス" to "kyanpasu")
- Rule Based Number Formatters, which allow users to convert numbers to their spelled out forms (i.e. 1,342 to "one thousand three hundred forty-two") and parse spelled out numbers into numeric values.

This tutorial will cover the basic ICU API for these two features utilizing built-in rulesets, and how to extend these features by writing new rulesets. These rulesets are written in a somewhat esoteric but powerful language, the syntax of which will be covered in this tutorial. In addition, the tutorial will go over how the preexisting rulesets are stored in the CLDR, and the process of submitting newly developed rulesets to the CLDR for wider usage.

12:30-13:30 - LUNCH

13:30-15:00 **SESSION 3 TUTORIALS**

Presenter:

Addison Phillips
Principal Globalization Architect, Amazon.com

Track 1 - Internationalization: An Introduction Part 1

What is internationalization? Culture and language are all around us and affect the way in which we expect our software to work--from the smallest app to the largest Web site. This tutorial describes the concepts behind internationalization, localization and globalization so you can start to build software that seamlessly responds to the needs of users, regardless of language, region, or culture. Start here to learn how to identify internationalization issues, develop a design, and deliver a global-ready solution, drawing on the presenter's wide experience.

- Identify Internationalization Issues
- Develop and Build a Global Product
- Work with Different Languages, such as Asian, Indic, and Right-to-left Scripts
- Deal with Cultural Issues that go Beyond Language

Presenter:

Tex Texin
Chief Globalization Architect, Xencraft

Track 2 - Web Internationalization

This tutorial, updated in 2019, 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.

Presenter:

Track 3 –

15:00-15:30 - Afternoon Refreshments

15:30-17:00	SESSION 4 TUTORIALS

Presenter:

Track 1 - Internationalization: An Introduction Part 2

Addison Phillips
Principal Globalization Architect, Amazon.com

What is internationalization? Culture and language are all around us and affect the way in which we expect our software to work--from the smallest app to the largest Web site. This tutorial describes the concepts behind internationalization, localization and globalization so you can start to build software that seamlessly responds to the needs of users, regardless of language, region, or culture. Start here to learn how to identify internationalization issues, develop a design, and deliver a global-ready solution, drawing on the presenter's wide experience.

- Identify Internationalization Issues
- Develop and Build a Global Product
- Work with Different Languages, such as Asian, Indic, and Right-to-left Scripts
- Deal with Cultural Issues that go Beyond Language

Presenter:

Track 2 – Put ICU to Work

Mihai Nita
I18N Senior Software Engineer, Google, Inc.

This tutorial gives an introduction to the Android´s internationalization and localization features, including a tutorial for developing an internationalized Android app from scratch (localizability, formatting, bidi, etc.)

Presenter:

Track 3 – Unicode Properties

Martin Dürst
Professor, Aoyama Gakuin University

Unicode uses a large number of properties to associate information to characters. Some properties provide very basic information, for example whether a character is a letter, a digit, a symbol or something else. Other properties are highly specialized, e.g. to drive Unicode normalization, the Unicode Bidirectionality algorithm, and other kinds of processing.

This tutorial will be of interest to intermediate and advanced users of Unicode. Topics covered will include an overview of properties and property classifications, how to obtain and extract property data (and metadata about properties) from Unicode data files or from programming language APIs, and how to use property data for various applications.

Thursday, October 17, 2019

09:00-09:15	***WELCOME & OPENING REMARKS***
09:15-10:00	**KEYNOTE PRESENTATION - Don't Believe a Word: Multilingual Typographic Systems and a 100-year Publishing Project**
Presenter: **Dr. Rathna Ramanathan** *Reader in Intercultural Communication and Dean of the School of Communication, Royal College of Art, London*	The Murty Classical Library of India aims to make accessible modern translations of Indian texts in print and online. In the first five years, 22 volumes in 12 different languages have been published. This keynote reflects on the delights and challenges of building a complex, multilingual typographic system for this unique 100-year publishing project as well as a subsequent research project which aims to create typographic guidelines for Indian languages and scripts.
10:00-10:30 - Morning Refreshments	
10:30-11:20	**SESSION 1**
Presenters: **Rafael Xavier de Souza,** *Senior Software Engineer, PayPal* **Arjun Madgavkar** *World Ready Software Engineer II, PayPal, Inc.*	**Track 1 - I18n Consistency, Speed, and Usability for JavaScript Apps** PayPal´s mission is to ensure that every person has the ability to participate fully in the global economy and without proper internationalization this would be impossible to achieve. In this session, you will hear from the engineers that designed the latest version of the JavaScript API that makes PayPal´s products world-ready. They will discuss both the high level ideas behind PayPal´s i18n ecosystem as well as the technical details regarding the following: • How to Build for React.js and Webpack Integration • How to Design Modular and Fine-grained Architecture that Eases Tree Shaking, Localization Abstraction, and Speed • How to Optimize for Size Using Static Code Analysis
Presenter: **Mark Davis** *Internationalization Architect, Google*	**Track 3 – The CLDR Tutorial** The Unicode CLDR project has become a well-established resource for system and application developers who need accurate information regarding culturally specific behaviors in their programs. This in-depth tutorial session explores the history and purpose of the CLDR project, types of data that are maintained, new features, and information about how individuals and organizations can leverage the large amount of data that is freely available through the CLDR project. Whether you are a first time adopter of CLDR, or have been using it for a long time, this session will highlight ways in which you can not only use CLDR, but also how you can contribute to it in order to provide the highest quality possible.

Presenters:

Anshuman Pandey
Natural Language Technologist, Script Encoding Initiative, UC Berkeley

Asmus Freytag
President, ASMUS, Inc.

Track 3 - Designing Internationalized Domain Names for the DNS Root Zone

Top-level domain names (TLDs) have come a long way from three-letter ASCII labels. Allowing Internationalized Domain Names (IDNs) in the Root Zone provides both opportunity and risks. IDNs require attention to issues that many users and developers are not familiar with and that go beyond what is addressed in specifications like IDNA2008 or UTS#46.

In 2013, ICANN started an effort to better understand the issues posed by IDN TLDs and to adopt a set of "Label Generation Rules" (LGRs) that would formally define the available IDNs for the Root Zone for each script. These LGRs not only define eligible characters, but also the permissible contexts that they may appear in, as well as equivalent character sequences (variants). The DNS Root Zone is global shared resource. The Root Zone LGR will simultaneously support up to 28 modern scripts, posing particular challenges, including cross-script issues.

This presentation highlights the issues and challenges presented by IDN TLD in general and the techniques used to mitigate them in an effort to balance the expansion of access to new writing systems with the security of the DNS Root Zone. Many of the techniques discussed would also be appropriate for more secure use of IDNs in other public zones or other identifier systems, such as user names. This also sets the stage for the following presentation topic: "Designing Domain Names for Indic Scripts". While that presentation is specifically about Indic scripts, it exemplifies many of the general issues presented here.

11:30-12:20

SESSION 2

Presenter:

Shane Carr
Senior Software Engineer, Google, Inc.

Track 1 – Internationalization in ECMAScript

JavaScript is a platform available not only in web browsers but also on mobile devices, IoT, cloud services, and many others. This means that building i18n features into JavaScript (formally ECMAScript) has a high impact on i18n library availability. ECMA 402 is a subcommittee of TC39 that standardizes features for the Intl object of ECMAScript. The subcommittee's work has previously included Intl.DateTimeFormat, Intl.NumberFormat, Intl.Collator, and others. This presentation will give an overview of ECMA 402 and the proposals currently underway, including Intl.Segmenter, Intl.Locale, date range formatting, unit formatting, compact decimal notation, and display names.

Presenters:

Rod Sheeter
Tech Lead/Manager, Google Fonts, Google, Inc.

Garrett Rieger
Senior Software Engineer, Google Inc.

Track 2 – Delivering Variable Fonts to the Web at Scale

Google Fonts delivers free, open source, fonts to billions of pages across the web. Recently we have significantly advanced our capabilities around delivering Variable Fonts (the latest font standards edition) and begun contributing to a new W3C standard (https://www.w3.org/Fonts/WG/webfonts-2018.html) for network delivery and reassembly of parts of font. We will discuss the core problems we face in delivering Variable Fonts at web scale, challenges around non-Latin fonts, what we are shipping, open source Variable Fonts, and advancements in the Open Source toolchain for dealing with fonts.

Presenter:

Asmus Freytag
President, ASMUS, Inc.

Track 3 - Designing Domain Names for Indic Scripts

Expanding on the topic of the preceding presentation, which covers Internationalized Top-level domain names (IDN TLDs) for the root zone in general, this presentation will focus on the particular issues faced by the family of NeoBrahmi scripts in use in India. The issues for Indic Scripts are sufficiently complex and detailed, and of interest in their own right given the relative lack of information for these scripts, that they deserve a focused presentation.

As part of an ICANN effort to better understand the issues posed by IDN TLDs and to adopt a set of "Label Generation Rules" (LGRs) that would formally define the available IDNs for the Root Zone for each script, including the Indic scripts. Such LGRs not only define eligible characters, but also the permissible contexts that they may appear in, as well as any equivalent character sequences (variants). The latest iteration of these Root Zone Label generation rules, RZ-LGR-3, added coverage for most of the Indic scripts.

This presentation highlights some of the issues and challenges specific to designing DNS for these scripts as well as the specific techniques and approaches used in mitigating the risk to the DNS. Many of the techniques discussed would also be appropriate for more secure use of IDNs for Indic scripts in other public zones or other identifier systems, such as user names.

12:30-13:30 - LUNCH

13:30-14:20

SESSION 3

Presenters:

Markus Scherer
Unicode Software Engineer, Google LLC

Steven Loomis
Software Engineer, IBM

Track 1 - ICU @ 20 Anniversary Edition - ICU Has Been Open Source for 20 Years!

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 code from Unicode, it provides cross-platform C/C++ and Java APIs.

This presentation will provide a brief history and overview of ICU, with emphasis on recent updates, including the latest support for Unicode 12.1 & Emoji 12 and the new Japanese era, better locale ID handling, better formatting output metadata, a new fine-grained mechanism for reducing the data size, and other changes. The presentation will also touch on ICU´s planned direction for future releases.

Presenter:

Alolita Sharma
Principal Technologist, AWS

Track 2 – Indian Language Support for a Billion+ Users in 2019 - A Survey of Challenges, Opportunities and Technology Solutions

Indian languages represent the cultures and history of more than 1.4 billion people. These languages also represent the fastest growth markets for mobile, AI driven platforms. This talk will present a snapshot of current language technology support on major platforms including web browsers, mobile devices and voice assistants. The talk will also discuss existing technical and UX gaps in input tools, fonts, conversational AI solutions as well as highlight solutions that are being developed by the top 5 technology global vendors.

Presenters:

Track 3 - Unencoded Scripts in the Unicode Standard: The 2019 Update

Anshuman Pandey
Natural Language Technologist, Script Encoding Initiative, UC Berkeley

Deborah Anderson
Researcher, Dept. of Linguistics, UC Berkeley

One of the core mandates of Unicode is to provide support for all of the scripts of the world. Although the encoding effort is challenging on several levels, the success is evident in each new version of Unicode. Four scripts were added in Unicode 12.0: two modern (Nyiakeng Puachue Hmong and Wancho) and two historic scripts (Elymaic and Nandinagari). Other scripts are in the pipeline for future publication, such as the three historic scripts Chorasmian, Dives Akuru, and Khitan Small Script.

Overall, the progress in encoding scripts has slowed from the high point in Unicode 7.0, which added 23 new scripts. However, there is active work being conducted behind the scenes by several individuals and the Script Ad Hoc Committee of the UTC, which assesses proposals and makes recommendations to the UTC through published reports.

For several of the scripts that remain unencoded, substantial research is required, such as for Old Uyghur, Mayan hieroglyphs, and a major extension of Egyptian hieroglyphs. Also, some scripts require meetings with users, in order to work through issues and receive approval from experts, before the proposals can be finalized (such as Cypro-Minoan and Leke). The total number of unencoded scripts remains large, with a slowing growing tally of 160+. This presentation will highlight some of these scripts and the challenges in their encoding. It will also foster discussion about ways to tackle the number of unencoded scripts within the next decade.

14:30-15:20

SESSION 4

Presenter:

Track 1 - Managing Your ICU Locale Data

Shane Carr
Senior Software Engineer, Google, Inc.

ICU4C binary distributions and ICU4J jar files have a large footprint, exceeding 11 MiB compressed size, largely due to the footprint of locale data. While this is usually fine for server-side applications, the increased overhead for application downloads and binary distributions are problematic for many users doing client-side i18n on smartphones and IoT devices. The data footprint of ICU is often a blocker for users adopting ICU, resulting in those products having subpar i18n support. ICU 64 introduces the ICU Data Build Tool, an approach for optimizing high-footprint data bundles to include only the locales and features users need in their ICU build. This session will be a deep-dive into how ICU locale data works and how to leverage the ICU Data Build Tool to reduce your locale data footprint.

Presenters:

Track 2 - Meet Shakespeare, the Netflix International Copy Testing Framework

Tim Brandall
Internationalization Manager, Netflix

Shawn Xu
Internationalization Program Manager, Netflix

Pu Chen

Words matter, and a simple phrasing or terminology change has more power than you would imagine when it comes to moving the needle on the number of people signing up for Netflix. We test multiple different versions of copy across many different languages in our product UIs and messaging, almost constantly, but these tests are time consuming to setup and read.

Enter Shakespeare. We set out to design a copy testing framework that would allow Product Managers, Linguists, and Copy Writers to test multiple different versions of the same source copy, in any language, on any platform, at the push of a button. Shakespeare empowers them to do just that, handling the A/B test allocation of the various versions of copy, then interpreting the results using machine learning, ultimately promoting the

Internationalization Engineer, Netflix

winning copy into production.

We´re looking forward to showing you how it was built, and we´ll also feature a demo.

It´s not every day you get to meet Shakespeare.

Presenter:

Andrew Glass
Senior Program Manager, Microsoft

Track 3 - Egyptian Hieroglyphic Sign Blocks in OpenType

In March 2019, Unicode 12 added nine format controls for Egyptian Hieroglyphs. These controls define how individual signs are arranged in the sign blocks that are typical of Egyptian Hieroglyphs and are therefore necessary for the faithful representation of this writing system in plain text. A prerequisite to adding these controls to Unicode was the ability to show that OpenType technology could handle the layout requirements they enable so that arbitrary structures can be formed. This turns out to be an interesting problem. This presentation will show how arbitrary sign blocks can be formed for Egyptian Hieroglyphs using existing OpenType technology; how they should be built up while typing Egyptian with an Egyptian IME; and how the overall solution can be extended to other Hieroglyphic scripts, particularly Mayan.

15:20-15:50 - Afternoon Refreshments

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

Presenter:

Martin Dürst
Professor, Aoyama Gakuin University

Track 1 - A Domain-specific Language for Unicode Properties

Unicode provides a large number of character properties in a variety of data formats. Every time this data is used, code has to be written again. Although the Unicode Consortium tries to use similar data formats wherever possible, differences have accumulated over time and backward-incompatible changes are strongly discouraged.

This presentation introduces a domain-specific language (DSL) for handling Unicode property data. The language allows to indicate property metadata (e.g. whether a property has binary or numeric values), to associate properties with data files, and to select subsets of properties or property values. The DSL is written in the programming language Ruby, and therefore easily allows to tie in additional processing steps at various stages. We will show various applications of this DSL, from quick checks to the creation of compact data structures. We will also present ideas of how metadata about properties might be published in a more streamlined way.

Presenter:

Igor Afanasyev
Director of Localization, Evernote

Track 2 - Localization Automation at its Extreme

In the past few years the topic of localization automation becomes more popular within the industry. Many companies offer APIs, connectors and command-line tools to facilitate automation. Many solutions are advertised as offering continuous localization. But if you dig deeper, you can see that these solutions are either not entirely automated, or require extra integration effort or constant maintenance on the client’s side, and often are built on top of 20-30-year old “conventional” localization process. In his talk, Igor will show how to rethink your localization infrastructure and pick the tools that unlock true end-to-end continuous localization: the one that requires almost no maintenance and is as lean and simple as it can possibly be.

Presenter:

Moriel Schottlender
Senior Software Engineer, Wikimedia Foundation

Track 3 - It´s All Backwards: How the Human Element Makes Supporting BiDi Difficult (and What to do About It)

One of the bigger and more confusing internationalization challenges we face is that of Right-to-Left support, and even more so the support of bidirectionality: when a piece of text mixes LTR and RTL in the same sentence. While the Bidirectional Algorithm gives us a great, solid support system for such situations, it cannot analyze and understand the meaning of individual pieces of text; human beings must make sure that when they are dealing with bidirectionality, they utilize the bidirectional algorithm support intentionally and correctly. As always when depending on the human element, this leads to common mistakes and mishaps in bidirectional support. Some of those are so common, even right-to-left speakers often cannot recognize they are wrong. Some mistakes are so deeply embedded, they have impacted social behavior and even led to lawsuits.

This session will go over examples of real life misuse and mistakes that happen when the bidirectional algorithm is not applied properly, will discuss how prevalent it is in society considering the majority of people are not language or computer savvy, and will give some pieces of advice on how to tackle the more common problems.

16:50-17:40 SESSION 6

Presenter:

Mark Davis
Internationalization Architect, Google

Track 1 – Emoji in Code

Fundamentally, emoji are ordinary Unicode characters. However, they exhibit some of the characteristics of more complex scripts, and therefore call for special attention to avoid problems. While many products do an excellent job at supporting emoji, others have problems. My goal here is to cover best practices for implementing emoji and testing software, so that you can identify the problems and fix them.

Good emoji support has surprising side benefits. It has led to better support for regular human languages: an improvement in Unicode handling in general, and more timely support of new versions of Unicode in operating systems and applications. For example, some applications, such as the database MySQL, didn´t handle characters with Unicode numbers above 65,535 (where most emoji are). By handling emoji, they´ve also enabled support for the full range of Unicode characters, such as less-common Chinese/Japanese characters, and characters in less common scripts.

Presenters:

Craig Cummings
Staff Consultant/Evangelist, VMware

Demin Yan
Senior Engineering Manager, VMware

Jessiely Juachon
Staff Engineer, VMWare

Track 2 - Total Hands-off, Fully Automated Product Globalization: Part 2 -- The Juicy Bits

Last year, VMware presented on its globalization microservices technology. It sparked some great questions about standardization, security, performance, high-availability, and more. This year, with this technology becoming open source, we will answer lingering questions, dig deeper into the technology, and share consumer successes. Those who know my love of live demos will not be disappointed as this time around we´ll showcase more features of this technology.

Additionally, we´ll also take a more in-depth look at the also-open-sourced, SonarQube static code analysis rulesets. We´ll talk about how to make the rulesets work in your environment, how to interpret their results, and how best to leverage this hands-off globalization realm.

Presenters:

Thomas Milo
Decotype,

Alicia González Martínez
Universität Hamburg

Track 3 - YaKabikaj: A Search Algorithm for Arabic-scripted Languages. Real-time Unicode Normalization

Unicode encoding of Arabic script is marked by regional diversity and typographic variation. On the one hand, a grapheme may have more than one code point. This happens when the grapheme has multiple superficial shapes according to different regional or historical traditions. An example of this is ك ARABIC LETTER KAF ك in contrast with the Persian equivalent, ک ARABIC LETTER KEHEH ک: both represent the exact same grapheme. On the other hand, typographic oddities sometimes receive special encoding. For instance, the basmala, i.e., the Arabic phrase bi-smi llāhi rraḥmāni r-raḥīm بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ, is included as a single code point, ﷽ ARABIC LIGATURE BISMILLAH AR-RAHMAN AR-RAHEEM ﷽. Similarly, many grapheme clusters (in the linguistic sense) can be encoded as single code points or decomposed into several code points. For example, آ ARABIC LETTER ALEF WITH MADDA ABOVE آ can be alternatively encoded as ا ARABIC LETTER ALEF ا plus ARABIC MADDAH ABOVE ◌ٓ. Other graphemic clusters cannot be decomposed: e.g., a grapheme cluster consisting of a denticle with a dot above or a dot below is encoded as a typographic ligature (0646 ARABIC LETTER NOON and 0628 ARABIC LETTER BEH respectively). In addition to the variation and ambiguity on the part of Unicode, Arabic script itself is characterized by the presence of optional diacritics, including the disambiguating dot patterns. This means that the actual shape of a given text may change extensively depending on the presence or absence of all the possible diacritics within the text. Furthermore, Arabic script exhibits a variety of orthographic customs that allow two or more graphemes to be interchangeable in some contexts. For instance, ة ARABIC LETTER TEH MARBUTA ة may be typically written as ه ARABIC LETTER HEH ه in many varieties of the Arabic language and Arabic scripted-languages. Last but not least, even the more Arabic-aware search algorithms, that allow for inclusion or exclusion of diacritics, cannot find all relevant matches when, i.e., the presence of only one particular diacritic is relevant. All these phenomena make string-search algorithms underperform when run on Arabic scripted-languages. To solve this problem, we present YaKabikaj, a search algorithm that performs a real-time normalisation of Arabic script.

18:00-19:00 - CONFERENCE RECEPTION

Friday, October 18, 2019

09:00-09:50	SESSION 7

Presenters:

John Watson
Software Engineer, Facebook

Casey Charlton
Program Manager, Facebook

Track 1 - FBT - The Heart of Facebook´s Localization Process

Facebook´s localization process is based on the FBT (Facebook Translation) concept and allows Facebook to ship localized products in dozens of languages on a daily basis. The FBT incorporates contextual and project information to the original string, and relies on CLDR data to intelligently handle variants such as those for gender and number. Recently, one of the authors, John Watson, brought the magic of JavaScript FBT markup to open source. Casey Charlton leads the internal use of FBT to drive a high-volume, rapid delivery localization process. This talk will cover the entire FBT lifecycle, from the developer through to the end-user.

Presenter:

Track 2 - When a Merperson is a Merman

Jennifer Daniel
Creative Director, Emoji, Google

As of Unicode 12.0, 64 emojis are defined as "gender inclusive" yet most platform renderings do not have unique designs to support the vast majority of these codepoints which do not specify gender. The only explicitly gender inclusive emojis all platforms support with unique designs are for Child (U+1F9D2), Person (U+1F9D1), Older Person (U+1F9D3).

Currently, to support gender inclusive codepoints (Gender inclusive can be defined as male/female to an equal degree, can neither be confidently identified as male/female, etc.) all major platforms default to a male or a female design. This results in some problems when someone texts their friend from a Microsoft device, "Love a good mansplain (U+1F926)" which presents as FEMALE and their friend, if reading from an iPhone will see, "Love a good mansplain (U+1F926)" which presents as MALE. So, even though both of these emojis map to U+1F926 they are presenting different genders. This creates cross platform inconsistencies and in some cases reinforces stereotypes.

Giving "ungendered" emojis a gender inclusive appearance is not trivial. Are there signifiers that can provide clues which don´t rely on a gender binary (haircut, clothing, color, body language, other facial features)? What level of detail is too much or not enough? Emojis demand an instant read so what markers will be effective at communicating a spectrum of gender presentations at a small size?

When exploring signifiers that have potential to communicate if a design is inclusive it is important to remember that gender lives dynamically on a spectrum and there is no single visual design "solution". However, we cannot ignore gender and Unicode continues to support existing and possibly future codepoints. Gender inclusive designs are intended to represent a person of any gender. You could argue that a gender inclusive design 'works' if the gender inclusive artwork is truly ambiguous to the viewer. But, due to each individual´s own perception or ascribing of gender, it´s possible people look at the existing "male" emoji designs and see "woman with short hair" or look at the "female" emoji designs and see "man with long hair". Given how fluid and dynamic language and gender presentation can be it´s not inconceivable that regardless of haircut, shirt color, accessories and other clues, gender inclusive emojis could be repurposed in some other manner and bring us somewhere else entirely.

Presenter:

Track 3 -

10:00-10:50	SESSION 8

Presenter:

Nova Patch
Director, Internationalization & Localization, Shutterstock

Track 1 - Locale Negotiation, Selection, and Persistence

Strategies to determine the best available locale for new users, control locale selection for opinionated users, and persist consistent locale usage throughout platforms and the end-to-end user experience.

- Negotiation: usage of Accept-Language header, GeoIP country, user input, supported locales, and the Unicode CLDR
- Selection: UX, best practices, and even more CLDR!
- Persistence: among different platforms and communication methods

---

Presenter:

Jennifer 8. Lee
Co-Founder, Emojination

Track 2 - Hanmoji: A Comparison of Chinese Characters and Emoji

Even though their dates of origin are millenia apart, the languages of Chinese and emoji share similarities the average smartphone user might find surprising. The Chinese character 火, for instance, visually resembles the 🔥 emoji and literally translated means "fire".

These "hanmoji" parallels offer a unique window into the evolution of Chinese Han characters and, by extension, the present state and future path of today's global emoji phenomenon. Many cultural assumptions have become hard coded into the language, The word for "good", 好, is the combination of 女 (woman) and 子 (son), or 👩👶.

In additional to showing some delightful mashups, the Hanmoji research also delves in the 214 Chinese Kangxi radicals as a technique to find the semantic gaps in modern emoji. In finding "direct", "indirect" and "associative" emoji equivalents, the comparison reveals the strengths and dearth of written Chinese and emoji. The presentation takes a deep dive into Hanmoji to guide new and old Unicode hands through Chinese lessons with a fresh, modern twist.

Presenter:

Dr. Ken Lunde
Type Architect, Chez Lunde

Track 3 - The History of Japan's Era Name Square Ligatures

2019 represents a very important year for Japan: a new era named Reiwa (令和) began on 2019-05-01. Japan's current and four previous era names are composed of two kanji (aka ideographs). Japan's Era Name Law (1979) explicitly requires this, and further states that the two kanji must be easy to read and easy to write, among other criteria. At some point in the late 1970s or early 1980s, a Japanese company coined so-called "square ligature" forms of the current era name at the time, Shōwa (昭和), along with the two previous ones, Taishō (大正) and Meiji (明治): ㍾, ㍽, ㍼. These characters became broadly supported in fonts from the time, and when the Heisei (平成) era began in early 1989, it subsequently received the same two-kanji square ligature treatment: ㍻. These four two-kanji square ligatures were included in the very first version of Unicode, Version 1.0 (1991), which set a strong precedent for giving Reiwa the same treatment, whose two-kanji square ligature was included in Unicode Version 12.1 that was released earlier this year. This presentation will explore the history and development of these five two-kanji square ligatures.We will show examples of how Zawgyi impacts Google products and describe approaches for processing Zawgyi text, including detection and conversion to standard Unicode via open-source libraries. We invite discussion of

how the technical community can best serve users of both the Burmese language as well as users of other languages that share the Myanmar script (such as Mon, Shan and Karen).

10:50-11:10 - Morning Refreshments

11:10-12:00	SESSION 9

Presenters:

Tex Texin
Chief Globalization Architect, Xencraft

Craig Cummings
Staff Consultant/ Evangelist, VMware

Mike McKenna
Globalization Strategist, PayPal, Inc.

Track 1 - Architecture Tradeoffs for Global Software Design

Presented by software internationalization experts, this session describes the typical architectural tradeoffs confronting developers building Unicode-based software applications.

Attendees will come away with an overview of many of the design decisions that must be made, possible solutions, including best practices, potential pitfalls, and criteria for choosing solutions. This is the first time this session is being presented, and the intent is to eventually expand the content to a comprehensive tutorial.

The range of topics includes:

- Usability considerations for end users and for developers, including practices for multiple form factors, formats, layouts and styles
- I18n API design considerations including data types based on standards and those lacking standards, public and commercial libraries, and modular design
- Processes for accepting, tailoring, overriding, updating and reviewing with stakeholders, sources of embedded standard data including: Time zone data, CLDR, ICU, locales, collations and Unicode
- Considerations for automation including metadata updates, and deployments
- The ins and outs of working with Encodings, Normalization, and Databases. For example, using NFKD in natural language processing to reduce size, increase speed, and increase value of results

---

Presenter:

Esko Clarke Sario
Consultant, Kotoistus / Oy DataCult Ab

Track 2 - Challenges of Localizing Emoji Character Names

An emoji is not always an accurate depiction of the object or feeling that it is named after, but surprisingly often it appears as a confusing image both to the sender and to the receiver. The localizer is faced with the choice of simply translating the English emoji names and keywords, or trying to read and interpret the image the same way as the users from her region and culture might do. In the presentation, we´ll discuss the pros and cons of each approach. To unexpectedly receive a Slightly Frowning Face in an otherwise happy message is not uncommon, even if you expected a smiling. The sender might have randomly chosen what looked like a happy face to her, in her device, while the receiver with a different device or app also have relied on the given keywords, and if the two friends used different languages in their devices, there´s a risk that the keywords don´t match cross the language border.

The artistic freedom in picturing an emoji gives us enjoyable pictographs with wide cultural and stylistic spectrum, but sometimes it contributes to breaking the communication instead of supporting it. The Unicode

Emoji Chart doesn´t list the required visual elements in detail, so the variation can be significant. Sometimes cultural and regional references are the reason for the different details, as there naturally is a need to make the image culturally relevant to the users.

In localization, while looking at the same image that many of the users will be shown, there´s a temptation to describe what you see, based on your cultural background. After all, the image is the decider for most users in their selection of emoji. Their reading of it relies on the visual clues included in the image, some of which are part of the emoji definition, others decided by the style, or added for regional or cultural relevance. However, the image is likely to change over time and from one app and system to another, and it is not feasible or acceptable to keep changing our localized names and keywords accordingly. The opposite alternative for the localizer would be to use the English emoji name and keywords as their source, exclusively. Maybe the solution can be found somewhere in-between?

Presenter:

Alexey Pychenkov
Software Engineer, Facebook

Track 3 - **A Hard Path from Zawgyi to Unicode**

How we deal with Zawgyi at Facebook: from our strategy for Myanmar to what we did so far (perf optimizations and how we use client font detector).

12:00-13:00 - LUNCH

13:00-13:50	SESSION 10

Presenter:

Claudia Galván
Early Stage Innovation

Track 1 - GDPR and Privacy Around the World

The General Data Protection Regulation (GDPR) has raised the bar on privacy in Europe, and new privacy laws are in effecting world. Protecting privacy is a continuous process and affects small and large companies alike. The process to manage privacy around the world impacts software development as well as business processes. This talk will provide a high-level overview of managing privacy in global products.

- Privacy today
- GDPR and other regulations
- Do’s and Don’ts on implementing privacy in your products
- Impact on business processes

Presenter:

Alolita Sharma
Principal Technologist, AWS

Track 2 – Growing Language Data in Unicode CLDR for the IoT Generation

The Common Locale Data Repository (CLDR) is Unicode´s most popular open data projects. CLDR is used by every organization on the planet that has web, mobile and IoT applications serving a global user base. Today CLDR contains locale specific formatting and parsing data for dates, time, currency, timezones, calendars. CLDR also provides crowdsourced translation and transliteration data. CLDR recently added character names and keyword data for Emoji characters too. This talk will discuss gaps in current data for existing locales such as address and name data as well as missing coverage for long-tail languages valuable in emerging

markets. The talk will also discuss suggested solutions to increase CLDR data coverage to support cloud and IoT applications.

Presenter:

Lucas Welti
World Ready Engineering, PayPal, Inc.

Track 3 - Time Zones and Real Life - A Customer Usability Case Study

When letting someone in another region know when something is going to arrive or when an event will occur, you need to know the time zone of the destination. There are a number of ways to give a user a choice of time zones. Most are based on the IANA (Internet Assigned Names Authority) list of time zones that originated from the Olsen Time Zone database. Using CLDR, some standard formats and variations are provided which have been implemented into most i18n libraries.

But is it intuitive when the sender is not familiar with the receiver's region? Will the time zone name fit in a usable mobile interface, when combined with the time? Is there an intuitive way to help the user choose which time zone to use for themselves or a desired location they are sending to?

This session explores how time zones are provided through CLDR, what the different presentation formats look like and what the benefits and problems are when using the various layouts. We will then take a look at how to handle formatting when a developer asks for a format that is not supported through CLDR. And finally, to help make an interface more intuitive, we will present methods to automatically choose a single time zone or list of time zones based on phone numbers or destination.

13:50-14:40 SESSION 11

Presenters:

Neha Utkur
World Ready Software Engineer, PayPal, Inc.

Sunjay Koshy
Full Stack Engineer, Uber Freight

Mike McKenna
Globalization Strategist, PayPal, Inc.

Track 1 - Natural Language Processing for High Confidence Multilingual Gibberish Name Detection

Payments and money handling organizations are very concerned about if people transacting in their systems are real people or not. This is closely related to efforts to reduce money laundering and terrorist funding. Neha Utkur and Sanjay Koshy have been working on a project to increase the confidence in real name detection. They have expanded on previous efforts by using Unicode character properties and normalization forms to decrease matrix sizes and increase speed in three character Markov chain statistical analysis of names, and have added a four character bit-map approach to do fast elimination of names that contain patterns not found in any real name. Their training data has come from multiple, openly available, government, health statistics, genealogical, and onomastic sources. They will present the problem and why original logic had to change, their approach to solve the problem, and results when tested against actual live data feeds filled with hackers trying to open fake user accounts.

Moderators:

Steven Loomis
Senior Software Engineer, IBM

Kristi Lee
Senior Program Manager,

Track 2 - CLDR PANEL: Most Impactful Learning from Onboarding CLDR

The Unicode Common Locale Data Repository (CLDR) provides key building blocks for software supporting the world´s languages. CLDR provides language and region specific locale data and structure for software internationalization. Companies and organizations collaborate to establish the industry-standard locale data in CLDR and it´s used broadly across platforms, open source libraries such as Unicode-ICU project, and used by many companies around the globe. In this session, meet some of the key users and contributors to CLDR

Microsoft

and hear about their learnings, pain points, and how to overcome those hurdles.

Panelists:

Craig Cummings
Staff Consultant/ Evangelist, VMware

Mark Davis
Internationalization Architect, Google

Jeff Genovy
Software Engineer, Microsoft

Presenter:

David Mohr
International Quality Engineer, Adobe and Adjunct Professor, Middlebury Institute of International Studies

Track 3 - Lost in Translation: Designing for Language AND Culture

Products are often designed for international customers and it is common knowledge that products are localized with one or more specific regions in mind. What often isn't as well known is that matching the language and geography is merely the starting point for this work as most regions also have specific market expectations. Some are obvious and well-documented but many are subtle, non-explicit cultural norms which must also be observed. Attention to international markets and customer cultures can happen at any point during the product development cycle, but when it happens could make or break the product's launch and ultimate success. Ideally, planning for these potential pitfalls should happen during the design process, so as to maximize cultural fit. Realistically, this often doesn't happen, but if at least marketing doesn't take heed, the results can be disastrous. We will explore many of the well-known examples of failure, looking at numerous case studies, some of which crippled the product's firm. We will also break out the different kinds of mistakes, as some are FAR easier to avoid than others...!

In a course on International Design that David teaches, the first lecture introduces the problem by surveying the failures at other companies with their products. He leads through the three general kinds of failures -- linguistic/localization, regional standards/internationalization, and then cultural-mismatch of the design, with both techniques to limit exposure as well as different definitions of "culture" to think about.

14:50 – 15:10 - Afternoon Refreshments

15:10 - 16:00	SESSION 12

Presenter:

Robert Cameron
Professor, Simon Fraser University

Track 1 - Efficient Search and Extraction from Compressed Text Using ZTF-8

In Big Data applications, large textbases are often stored in compressed form using well-known general purpose dictionary-based encodings such as LZ4. While reducing storing requirements, these compressed forms generally require the time-consuming process of full decompression to support search and extraction of relevant text data. To address these concerns, ZTF-8 is designed as a dictionary-based compression format for Unicode documents with the ability many forms of search and text extraction operations without

full decompression. ZTF-8 is also designed for efficient high-performance decompression using the parallel methods of the Parabix framework.

Moderators:

Markus Scherer
Unicode Software Engineer, Google LLC

Steven Loomis
Senior Software Engineer, IBM

Track 2 - ICU PANEL: ICU User's Panel

With growing number of people coming on to the Internet in India, consumption and creation of content in Indic Languages is on the rise. There is a wide gap between demand for content and its availability, less than 0.1% of digital content on the Internet is in languages other than English.

Transliteration helps in bridging this gap for many utilitarian cases.

In a geography like India, where many languages are used, many business needs are solved through transliteration and not a real time machine translation. With literacy rates at above 70% in the country most businesses run in native languages and people find it easy and efficient to understand and use their own language. This is also corroborated by the fact that the local language daily circulations are ten times greater than English in the entire country.

Transliteration is a critical need for phonetic languages like Indic. Issues and factors that influence the process of transliteration relate to the nature of the script, language and the encoding used. In every language, the same words may be written in alternate ways. But, when it comes to transliterating between a non-phonetic language like English into a phonetic language like Hindi or Tamil, the problem is non-trivial. While there are several approaches including variants of rules based, statistical, machine learning and so on, attention to the nature of the issues and factors that affect the quality of transliteration, are vital to the preparation and curation of data. It also helps in making a choice for the process one may want to use or implement.

The paper deals with transliteration between two phonetic languages, transliteration between a non-phonetic language (English in case) and phonetic languages and vice versa. Hindi (An Aryan language) and Tamil (A Dravidian Language) languages are taken as examples to outline specifics. There are also comparison of some select processes with respect to the performance, efficiency and accuracy of transliteration.

Presenter:

Igor Viarheichyk
Engineering Manager, Samsung

Track 3 - Internationalization in Advanced Driver Assistance Systems

Advanced Driver Assistance Systems, or ADAS, is a rapidly emerging technology that is often associated with embedded systems, robotics, and AI, but rarely with internationalization. In this session, we discuss the importance of internationalization in ADAS platforms, practical approaches of adding i18n support into ADAS middleware, and unique challenges in this area.

16:10 - 17:00	CLOSING SESSION

Moderators:

Martin Dürst
Professor, Aoyama Gakuin University

Alolita Sharma
Principal Technologist, AWS

Lightning Talks

This traditional closing session will be a series of lightning talks of 5-10 minutes each, followed by extremely short closing remarks. The talks should be related to internationalization, localization, or any other of the topic areas listed in the Call for Participation. 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, or a conclusion or question you are taking home from the conference.

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 session.