CONFERENCE PROGRAM

Monday, September 10, 2018

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:

Craig Cummings
Staff Consultant/ Evangelist, VMware

Mike McKenna
Globalization Strategist, PayPal

Tex Texin
Chief Globalization Architect, Xencraft

Track 2: Introduction to Unicode and Beyond

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

Presenter:

Norbert Lindenberg
Internationalization Solutions Developer, Lindenberg Software LLC

Track 3: Creating Fonts for Brahmic Scripts

Brahmic scripts such as Devanagari, Tamil, Thai, and Burmese have complex requirements for mapping characters into correct arrangements of glyphs, including glyph reordering, conjunct formation, mark stacking, and mark positioning. OpenType, the font technology supported in all major operating systems, and Apple Advanced Typography, the technology in Apple’s operating systems, use different approaches to support the creation of fonts that meet these requirements: OpenType provides specialized shaping engines for several scripts as well as the universal engine for numerous others, while AAT provides a flexible generic shaping language. This tutorial discusses the requirements of Brahmic scripts and the approaches used by the two technologies, and provides practical guidelines for creating fonts.

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* **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.
Presenters: **Steven Loomis** *Software Engineer, IBM* **Elnaz Sarbar** *Program Manager, Google, Inc.*	**Track 3: CLDR/Survey Tool: How to Add New Locales and Add Locale Data Using the CLDR Survey Tool and XML for Seeding New Locales** How to add new locales and add locale data using the CLDR Survey Tool and XML for seeding new locales.

12:30-13:30 - LUNCH	
13:30-15:00	**SESSION 3 TUTORIALS**
Presenter:	**Track 1 - Internationalization: An Introduction Part 1**
Addison Phillips *Principal Globalization Architect, Amazon*	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.
Presenter:	**Track 2 - Web Internationalization**
Tex Texin *Chief Globalization Architect, Xencraft*	This tutorial, updated in 2018, 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 - Introduction to Android's Internationalization**
Mihai Nita *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.)
15:00-15:30 - Afternoon Refreshments	
15:30-17:00	**SESSION 4 TUTORIALS**
Presenter:	**Track 1 - Internationalization: An Introduction Part 1**
Addison Phillips *Principal Globalization Architect, Amazon*	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.

Presenters:

Steven Loomis
Software Engineer, IBM

Shane Carr
Software Engineer, Google, Inc.

Track 2 – Put 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, and is now hosted by Unicode itself. 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 (including the new fluent API), 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, integrating ICU into an application's development process, and how to get involved in the ICU development community.

Presenter:

Behnam Esfahbod
Software Engineer, Quora, Inc.

Track 3 - Get Your Head Around Bidirectionality!

We know when the software is broken for right-to-left languages like Arabic, Persian, or Hebrew, but often the solution is either not clear, or fixing it with out-of-place patches won't worth the costs down the road. Like other areas of i18n, bidirectional layout and right-to-left language support need deliberate design in the user-interface stack, and without good architecture it won't be useful for the developers or the users.

In this tutorial, we first learn how to think in right-to-left and how it mirrors into left-to-right directionality. We then look at the common problems in bidirectional applications and how to address them with generic solutions and standard algorithms.

This tutorial is suitable for anyone not familiar with right-to-left languages or bidirectional design, or interested to learn how to develop solutions for this area.

Tuesday, September 11, 2018

09:00-09:15	***WELCOME & OPENING REMARKS***
09:15-10:00	**KEYNOTE PRESENTATION - The Advent of Mayan Script Encoding: Mapping the Last Frontiers of Mayan Hieroglyphic Decipherment**
Presenter: **Carlos Pallan Gayol** *Archaeologist and Epigrapher, Department of Old American Studies and Ethnology, University of Bonn*	Mayan hieroglyphs rank among the most visually complex writing systems ever created. Deciphering them has entailed a 200+ year scholarly quest, but this task is not yet completed and posits an inviting challenge for applying new tools from the information-age, culminating in Mayan script encoding. This keynote highlights the latest milestones attained in this pursuit by the NcodeX Project, where Carlos Pallan collaborates with Dr. Deborah Anderson at UC Berkeley, the Script Encoding Initiative and members of the Unicode advisory board. Stemming from research funded by Unicode's "Adopt-a-Character" Program, it has been possible to produce new database tools and advanced functionalities, capable of mapping and analyzing all the textual contents of the extant Mayan books or Codices by relying on a novel catalog of Mayan signs with assigned code points. Over 900 signs and variants have been mapped in the ancient texts while a new open-type Mayan font is being developed, as published during the last UTC Meeting celebrated at Google, and featured by the NEH in their magazine Humanities. Next steps will include expanding this research from the Codices into the vast realm of Mayan hieroglyphic inscriptions from the Classic period and creating an open-access text repository powered by the state-of-the-art Mayan-READ interface, in collaboration with an international team of researchers and developers in the US, Mexico and Europe.
10:00-10:30 - Morning Refreshments	
10:30-11:20	**SESSION 1**
Presenter: **Mike McKenna** *Globalization Strategist, PayPal*	**Track 1 - Global First: Web First? Mobile First? Who's First?** When creating software, application designers must make a number of choices. Among them is what language to use for the user interface? When onboarding new customers, we always make the Customer First. We allow them to choose their desired language and we persist that choice in a profile to use later for later email communications, reports, and sms messages. But what happens if they log in through their mobile device? A mobile device is more an appendage of the user's brain than a desktop or laptop computer is. If the device is in a different language than the user's profile, what language do you send to the device? How do you choose what formats to use for internationalized dynamic information? This talk will explore these questions and more, focusing on data analysis made over a number of months of user's actual activities and differences between devices used. It will walk through how that data can be used to make market decisions and reduce customer support issues while enhancing the user experience. Topic to

include:

- user experience flow and where i18n and l10n step in to craft the user experience
- actual user data and the differences by device
- decisions made as a result of user data
- how the device dictates the locale negotiation
- choosing locale for live session, for messaging, for sms
- designing software to make the right dynamic i18n decisions based on context

---

Presenters:

Markus Scherer
Unicode Software Engineer, Google, Inc.

Steven Loomis
Software Engineer, IBM

Track 2 - New in ICU and CLDR

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.

The Unicode Common Locale Data Repository (CLDR) is the industry-standard locale data project where companies and organizations collaborate on the data needed to support many languages in operating systems, libraries like ICU, keyboard apps, etc.

This presentation will provide a brief overview of ICU and CLDR, with emphasis on recent updates, including the latest support for Unicode 11.0 & Emoji 11.0, new keyboard mapping support, and other changes.

Presenters:

Anshuman Pandey
Natural Language Technologist, Script Encoding Initiative, University of Michigan

Deborah Anderson
Technical Director, Unicode Consortium and Researcher, Dept. of Linguistics, UC Berkeley

Track 3 - Undeciphered Scripts in the Unicode Age: Challenges for Encoding Early Writing Systems of the Near East

There are several important historical scripts that are not fully deciphered, such as the Proto-Sinatic alphabet and the Byblos syllabary. They are actively studied and their signs are discussed and depicted in scholarly materials and books on the history of the 'alphabet'. However, these scripts cannot be represented in digital content because they are not encoded in Unicode. One primary factor for the lack of an encoding is the status of decipherment. While the sign repertoire for several of such scripts is known, the absence of information about the meaning of various signs pose particular problems for script researchers and technical standards committees.

How to provide support for these scripts without knowing the meaning of all their signs? Which signs could be considered distinctive characters and which are variants? Most significantly, how to meet the needs of scholars and other users of such scripts in the age of Unicode?

This session will discuss these relevant issues from two perspectives:

- Anshuman Pandey will approach the topic from the viewpoint of a script researcher, with examples drawn from Early Alphabetic, also known as Proto-Sinaitic, which is descended from Egyptian Hieroglyphs and

the ancestor of the Phoenician script, and in fact, all alphabets, abjads, and abugidas. He will address issues such as: How to define such a writing system as a 'script' in Unicode? How to model scripts for which decipherment is ongoing? At what point can an undeciphered script be suitably proposed for encoding?

- Deborah Anderson will speak as a member of the Unicode Technical Committee (UTC) and the ISO subcommittee on character sets, and as head of the Unicode Script Ad Hoc, which reviews script proposals before they are sent to the UTC. From the viewpoint of these committees, how many proposals for partially deciphered scripts been approved? What questions have the committees raised on such proposals? This portion of the session will include examples of such proposals, such as Cypro-Minoan, and conclude with observations on relevant features.

11:30-12:20	**SESSION 2**
Presenter:	**Track 1 - TBA**
TBA	

Presenters:	**Track 2 – ICU and CLDR**
Markus Scherer *Unicode Software Engineer, Google, Inc.* **Steven Loomis** *Software Engineer, IBM*	ICU (the International Components for Unicode) is a widely-used implementation of Unicode, and CLDR (the Common Locale Data Repository) is a popular source for language and region-specific locale data. Both are needed by globalized applications. Now that ICU is also hosted by the Unicode Consortium, this session will discuss users of either - or both - of these foundational projects. The two projects will be introduced very briefly, and then continue with a panel discussion focused on the experience of direct consumers of (and contributors to) both projects. Topics discussed will include benefits and challenges of using ICU, how ICU consumes CLDR data, as well as tips and techniques for CLDR implementers.

Presenter:	**Track 3 - The Mongolian Script: What's Going On?**
Liang Hai *Multilingual Font Technician*	The Mongolian script has been encoded in Unicode for nearly two decades, but is still struggling today to gain basic day-to-day usage in the digital world. So what are the reasons behind the failure of implementation? How are experts trying to improve the situation? From this talk you will learn about: • A crash course on what is the Mongolian script and how it works (from phonetic letters to graphemes) • A brief history of the Mongolian encoding (from pre-Unicode encodings to the Unicode encoding, and what are not working in the latter) • Ongoing efforts (various groups of experts and various meetings, and how to participate) • Tough lessons learned from the Mongolian encoding • How Do We Support the Mongolian Script Better in the Future

12:30-13:30 - LUNCH

13:30-14:20	**SESSION 3**
Presenter:	**Track 1 - Slicing and Dicing Unicode Properties**
Martin Dürst *Professor, Aoyama Gakuin University.*	Unicode defines a large number of character properties for character classification and algorithms. This talk compares the two main ways of implementing property support, and discusses their advantages and disadvantages based on actual implementation experience. This talk is suited for users dealing with Unicode properties as well as for implementers. Examples of Unicode character properties range from general category (letter, number, symbol,...) to specialized properties for tasks such as bidirectional layout and normalization. Many programming languages, libraries, and regular expression engines provide access to these properties. We provide some insights and results based on experimental work on the Onigmo regular expression engine that is used by the programming language Ruby. Leaving special properties such as character name aside, the two main ways of implementing property support are inversion lists and folded tries. Moving from the former to the later, we succeeded to increase the number of properties covered from 62 to 76, and the number of property values covered from 554 to 1009, all while reducing the memory necessary from 240kB to 214kB. In addition, elimination of binary search made raw property checks up to 9 times faster, and lookup of specific property values from characters up to 65 times faster. Comparing the two methods from a conceptual viewpoint, it is interesting to observe that when arranging property values in a huge table with characters as rows and properties as columns, inversion lists look at this table one column a time, whereas folded tries look at it one row at a time. Folded tries take advantage of the fact that most property value combinations are shared by a large number of characters, whereas inversion lists take advantage of the fact that subsequent characters often share the same property value. This understanding can lead to further improvements. [Work done jointly with student, Takumi Koyama]
Panelists:	**Track 2 – PANEL: Preparing for May 1, 2019: Japan's New Era Name**
Mark Davis *Chief Internationalization Architect, Google Inc.* **Kristi Lee** *Senior Program Manager, Microsoft* **Steven Loomis** *Software Engineer, IBM* **Ken Lunde** *Senior Computer Scientist 2, Adobe Systems, Inc.*	Please join this panel discussion about Japan's new era name with representatives from Adobe, Google, IBM, Microsoft, and other leading companies. They will discuss the major challenges for the software migration needed by May 1.

Presenters:

Huidan Liu
Senior Software Engineer, Institute of Software of Chinese Academy of Sciences

Jian Wu
Professor, Institute of Software of Chinese Academy of Sciences

Bo An
Engineer, Institute of Software of Chinese Academy of Sciences

Track 3 - Introduction to the Chinese Character Repository Project

Although more than 80000 Chinese characters are included in the Unicode character set, we still see some rarely used characters which are out of Unicode in many situations. The Chinese character repository project aims to solving the problem thoroughly. Many experts attended the project to collect uncoded characters which are used at any time from ancient China to present China. As estimated, about 300 thousand Chinese characters, 100 thousand ancient Chinese characters and 100 thousand minority characters will be found. More than one million code points will be used to express those newly found characters which are not coded in Unicode at present. We introduce the procedure and technologies related to the project.

14:30-15:20

SESSION 4

Moderator:

Alolita Sharma
Board Director, Unicode Consortium

Panelists:

Yiying Lu
Founder and Creative Director, Yiying Lu Inc.

Riitta Koikkalainen
Information Specialist, Kotoistus/The National Library of Finland

Megan O'Neill
Senior UX Designer, PayPal

Track 1 - Panel: Achieving Cultural Representation and Diversity with Emoji

Is it possible to achieve cultural representation and diversity with Emoji?

Join in for a discussion talking with leading UI/UX designers who have submitted successful Unicode emoji proposals to increase diversity and global representation on devices and platforms the world uses. This will explore and cover topics including impact of emoji on digital culture, interaction design factors and implications, and driving user engagement with representative emoji.

Presenter:

Anshuman Pandey
Natural Language Technologist, Script Encoding Initiative, University of Michigan

Track 2 - Developing a Successful Proposal for Encoding a Script in Unicode

The Unicode standard encompasses more than one hundred scripts and associated data, and it continues to grow with the inclusion of additional scripts and characters. While many of the world's writing systems have been added to the standard, there are more than one hundred scripts that remain unencoded. When users seek information about a script in the standard, the first document that they encounter is the code chart. This document is a concise visualization of an encoding that presents the characters of a script, their code points and names, and their graphical depictions. But the code chart also tells a story that is often unheard. At one point, the script in the chart was unencoded. What were the practices and processes that led to its encoding? There is no single answer to this question. The chart, along with the core data and core specification, represents a complex process of standardization for which best practices, conventional processes, and packaged solutions do not exist. This presentation will provide insights into the practices of the script-encoding process and the decisions that inform the development of an encoding model, creation of a repertoire, and selection of representative glyphs. It will also discuss Unicode principles that guide the encoding of every script, such as the character-glyph model and unification. This talk will also provide details on identifying a script for encoding, analytical methods for understanding its encoding model and repertoire, engagement and collaboration with users, and the standards approval process itself.

These aspects of the script-encoding process will be highlighted using important, related historical scripts that have been recently encoded, approved, or that are pending approval, namely Old Sogdian, Sogdian, Elymaic, and Khwarezmian. These are scripts descended from Imperial Aramaic, which became differentiated from their common ancestor and developed into distinctive scripts. While they share similarities in structure and representation, they present different issues related to script encoding. The discussion of the script-encoding processing through the lens of these related Aramaic-based scripts will appeal to Unicode experts and users of all levels, as well as to those interested in developing proposals for encoding scripts or characters.

Presenter:

Ken Lunde
Senior Computer Scientist 2, Adobe Systems, Inc.

Track 3 - Ten Mincho – To Boldly Go Where No Japanese Font Has Gone Before

This presentation is intended to convey technical details about how the "Ten Mincho" typeface and its fonts, which is the latest Adobe Originals Japanese typeface design whose fonts were released at the end of 2017, were developed, and how they "boldly go where no Japanese font has gone before."

The many ways in which the Ten Mincho fonts are unique or different from conventional Japanese fonts will be explored, such as the glyph set and Unicode coverage, kanji repertoire, incredibly rich Latin support to include italics, an incredibly large number of OpenType features some of which are language-sensitive, Unicode variation sequences, and even color SVG glyphs. The deployment format, which includes a separate style-linked italic face, is a space-efficient OpenType Collection.

The many ways in which the Ten Mincho fonts are unique or different from conventional Japanese fonts will be explored, such as the glyph set and Unicode coverage, kanji repertoire, incredibly rich Latin support to include italics, an incredibly large number of OpenType features some of which are language-sensitive, Unicode variation sequences, and even color SVG glyphs. The deployment format, which includes a separate style-linked italic face, is a space-efficient OpenType Collection.

15:20-15:50 - Afternoon Refreshments

15:50-16:40	SESSION 5

Presenters:

Riitta Koikkalainen
Information Specialist, Kotoistus/The National Library of Finland

Esko Clarke Sario
Consultant, Kotoistus/Oy DataCult Ab

Track 1 - The Case of the Three Monkeys: Localizing Emoji Names and Keywords

What aspects should be considered in localization of emoji names and keywords? Each emoji has its original meaning, but in addition to this, there are a variety of other meanings that exist among users. As it appears, many emojis are based on a need that is widely geographically applicable, while others have arisen out of the needs within a specific area and culture. In both cases, the actual usage of the emojis can be inconsistent, and sometimes it even contradicts the original meaning. In localization, local communication, as well as intercultural one, needs to be accommodated. The challenges relate to the more general discussion of different cultural identities versus universality. The history of the symbol should be reflected keeping in mind that it also needs to be understood globally.

Presenter:

Norbert Lindenberg
Internationalization Solutions Developer, Lindenberg Software LLC

Track 2 - Integrating the Development of Encoding, Font, and Keyboard

Traditionally there have been long gaps between the encoding of a script in Unicode and the development of fully functional fonts and keyboards for it. In some cases, this has led to the development of non-conforming or incompatible shaping engines, fonts, and keyboards, Burmese Zawgyi being a well-known example. In other cases, gaps or mistakes remained in the specifications, such as incomplete documentation of valid syllables.

This talk presents ideas for enabling the concurrent development of the encoding and at least one font and keyboard for new scripts. It discusses technical enhancements that would make such development as well as testing with experts and normal users of the script more feasible. Audience participation is strongly encouraged.

Presenter:

Roderick Sheeter
Tech Lead/Manager of Google Fonts, Google, Inc.

Track 3 - CJK on Google Fonts

Google Fonts delivers free, open source, fonts to billions of pages across the web. In recent years we have improved our quality and added support for a range of languages but were never able to support Chinese, Japanese, or Korean. Recently we have explored, validated, and shipped a data-driven CSS-only solution, leveraging unicode-range and Google's web index data to intelligently partition fonts to minimize transfer size, number of requests, and latency and optimize cross-site caching when delivering Korean fonts. This solution will also apply to Japanese, Simplified Chinese, and Traditional Chinese.

16:50-17:40	SESSION 6

Presenter:

Rob Cameron
Professor, Simon Fraser University

Track 1 - What do Intel's New AVX-512 Instructions Mean for High-Performance Unicode?

Long gone are the days when processors were limited to working with 8-, 16- or even 32 bits of data at a time. Single instructions operating on multiple data elements (SIMD) in 128-bit registers have long been available on Intel, AMD, Power PC and ARM processors. With the widespread availability of AVX2 technology in Intel processors beginning 2013, SSE2 instructions were generally extended to operate on 256-bit registers. In the

past year, we have seen the introduction of Intel's new 512-bit registers and AVX-512 instructions in high-end workstations as well as servers, notably including Apple's new iMac Pro workstations. In this talk, we discuss these processor trends, how they affect the performance of Unicode processing software in general and how the high-performance Parabix regular expression engine scales up to take advantage of these new capabilities.

Presenter:

Track 2 - Top Issues in Universal Acceptance of Non-Latin Email Addresses and Domain Names

Jim DeLaHunt
Principal, Jim DeLaHunt & Associates

The next one billion internet users use a wide variety of languages and scripts. They will demand email addresses, and domain names, in scripts they can easily read. This challenges apps and systems to provide Universal Acceptance (UA) — of all domain names and email addresses, from http://普遍接受-测试。世界 to شينام @ أشوك. الهند to données@fußballplatz.technology. We explain the most troubling obstacles and the most inspiring successes in Universal Acceptance encountered by the Universal Acceptance Steering Group. From concern over cross-script confusables to major email platforms launching support of internationalized addresses, it has been an exciting year.

Presenters:

Track 3 - Analyzing Support for Text Layout on the Web

Addison Phillips
Principal Globalization Architect, Amazon

Richard Ishida,
Internationalization Lead, W3C

The primary mission of the W3C Internationalization work, and of the W3C itself, is to create a Web for All. A particular area of interest and focus for the W3C is styling the layout of content, in web pages and in digital publishing. Much of this can be addressed in CSS, but there are other technologies that also need to take such factors into account, such as Timed Text, WebVTT, SVG, XSL, and to some extent markup models such as HTML, etc. It is particularly concerned with the mechanics of text, such as rules for line-breaking & justification, local approaches to expressing emphasis or decorating text, localizing counter styles, supporting bidirectional text in markup, initial-letter styling, hyphenation, page layout, etc.

Recently the W3C has been making additional efforts to better understand the needs of the various writing systems and cultures around the world, and communicate those to specification and browser developers. This talk will look at some of the things that are currently in progress or beginning, as well as possible future directions.

18:00-19:00 - CONFERENCE RECEPTION

Wednesday, September 12, 2018

09:00-09:50	SESSION 7

Presenters:

Craig Cummings
Staff Consultant/ Evangelist, VMware

Mike Fang
Staff MTS, VMware

Demin Yan
Senior Engineering Manager, VMware

Edwin (Zhenhui) Yang
Senior Engineer, VMware

Track 1 - Total Hands-off, Fully Automated Product Globalization

With the advent of SaaS and Agile software development processes, companies like VMware must deliver web/cloud applications in an ever-fast paced environment – moving from traditional, long releases to releases as short as days or even daily. The challenge of pace represents significant challenges to efficiently deliver sim-ship localized products with a high level of quality. In this technical session, VMware will introduce several automation solutions, such as those listed below, that revolutionize the pace of software globalization to be practically 'hands-off':

- Highly accurate static code analysis tools with customized rules for globalization that find issues early in the development cycle and help determine scope.
- A cross-programming language, microservices framework that solves not only localization issues at scale, but internationalization issues as well.
- Automated tools for delivery and reception of translations within the framework of continuous development and integration.
- Tools that create as well as execute test scripts based on scope from other parts of automation such as static code analysis.

Presenter:

Zibi Braniecki
Senior Platform Engineer, Mozilla

Track 2 - Javascript Internationalization API 2018

TC39, the working group behind JavaScript programming language, is proud to present the 2018 edition of the language together with the 4th edition of ECMA402 - its Internationalization API.

The 2018 edition brings a number of new APIs ranging from an addition of the first major new API in years - 'Intl.PluralRules', but also rich formatting of the Intl output via `formatToParts` and completion of the unicode extension keys support. What's more important, the ECMA402 working group is close to finalizing a number of major APIs including 'Intl.ListFormat', 'Intl.RelativeTimeFormat', 'Intl.Segmenter' and a major revision of its 'Intl.NumberFormat'.

This session will report on the progress and then ask the audience for feedback on the work before finalizing the specification.

Presenter:

Fesseha Atlaw
Founder, Engineer, Dashen Engineering Company

Track 3 - A Unicode Success Story - Transforming Ethiopic/Amharic from an Endangered Script to a Thriving Dynamic User of Digitization

When Ethiopia was taken over by a military Communist dictatorship in 1974, in an effort to re-write history the new government embarked in destroying all ancient church and monarchy related documents and books written in the Ethiopic script. Ethiopic was labeled as an archaic script that needed to be replaced by Latin. The major argument was that Ethiopic has way too many characters to be computerized and the language itself was outdated by "modern standard" and that keeping the script will render the country to remain behind civilization. While studying & working in the United states as hardware engineer, I was forced to digitize Ethiopic to be able to fulfill my life long dreams of continuing to write Amharic books and plays. In early 1980's I started designing fonts in Ascii framework pixel by pixel and finding ways to accommodate over 375 Ethiopic Characters into a computer word processor. Ethiopic was incorporated into the Unicode in 1991. Ethiopic is now used by about 30 million computer users in and out of Ethiopia and Eritrea. Now there are over 1,000 applications specifically made for Ethiopic use, that number growing exponentially. This should be mainly credited to the Unicode initiative to include Ethiopic in the standard early on. The presentation will chronicle the development of Ethiopic form being an endangered script to one of the most dynamic and thriving script worldwide.

10:00-10:50 SESSION 8

Presenters:

Erkki Kolehmainen
Consultant, Kotoistus/Oy KREST Sales and Consulting Services Ltd.

Esko Clarke Sario
Consultant, Kotoistus/Oy DataCult Ab

Track 1 - Country Model for Initial CLDR Data Collection – a Finnish Case Study

When CLDR data needs to be populated for a location that is lacking most or all of the content, a national model of co-operation both could and should be considered.

In Finland, the development of the model began some ten years ago, as the languages of Finland were added into the CLDR database. From the onset, the work was done as an open national project with most of the contributions coming from volunteering specialists. However, anyone could contribute and take part in the discussion. The project co-ordination was initially and continues to be financed by the Ministry of Education and Culture.

The project was named “Kotoistus”, and it started with collecting elementary data for the Finnish language. This has been reviewed and updated continuously. The next languages that were reviewed were Swedish, as spoken in Finland, and the Sami languages of Finland. These languages will be focused on for upcoming versions of the CLDR.

As the years have passed, the initiative has grown to be more or less permanent process and organization. Openness is still of importance: any interested party may be involved in the process in order to reach best possible data quality. All parties' suggestions are reviewed by the steering committee, and, if there is relevance with good reasoning, the suggestions are promoted further. In addition to CLDR, recommendations based on the gathered data are given to service providers and vendors. In addition, the recommendations are available to the general public in clear text form.

With a national initiative of this kind, the smaller languages and locales can reach high data quality and a good level of completeness. Their local data becomes available for global localization, and their chances increase for well localized products and services. The local becomes visible in global context.

Presenter:

Alolita Sharma
Principal Technologist, Amazon Web Services (AWS)

Track 2 - What's New with GlobalizeJS?

GlobalizeJS is one of the most popular open source JavaScript internationalization libraries in use today. This library is leveraged both by large enterprises and by startups to support i18n and L10n. It interfaces with client platforms (e.g., via React) and server implementations (e.g., via NodeJS). GlobalizeJS uses Unicode CLDR data and closely follows the UTS#35 specification. This talk will introduce the key features of GlobalizeJS and then highlight new capabilities, performance optimizations and data distribution mechanisms that have been added recently. The talk will also cover feature requests yet to be implemented and how you can contribute.

Presenters:

Shane Carr
Software Engineer, Google, Inc.

Jeremy Hoffman
Software Engineer, Google, Inc.

Luke Swartz
Product Manager, Google, Inc.

Track 3 - Fixing Burmese: Dealing with Zawgyi

Many Burmese speakers use fonts and keyboards that generate a "font hack" encoding called Zawgyi. This non-standard use of Myanmar-script code points breaks text matching, sorting, spell checking, word wrapping, and other text processing.

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

Presenter:

Roozbeh Pournader
Internationalization Engineer, WhatsApp

Track 1 - Internationalization at WhatsApp and Facebook

Internationalization efforts at WhatsApp and other Facebook products will be introduced, as well as unique challenges and the lessons and techniques learned that are useful for app developers.

Presenters:

Robert Heinz
Globalization Product Manager, Nike, Inc.

Aaron Presley
Senior Software Engineer, Nike, Inc.

Track 2 - Simpler Internationalization in Modern React Applications

Scaling out internationalization best practices across a large number of engineering teams can be difficult. In this session, Aaron Presley will discuss partnering with Nike's community of front-end engineers for the creation of the Nike I18N module.

This module enables an engineer to internationalize their front-end experiences with minimal setup & configuration time. It has pseudo-localization support out of the box; makes it easy to import & export translatable strings to Nike's Localization Platform; and works seamlessly with projects using React, Facebook's popular JavaScript library.

By leveraging the native INTL object and working in conjunction with a targeted polyfill approach it embraces a lean approach to web internationalization while continuing to provide a performant & premium user experience. To wrap-up the session, we will take a look at how it compares to other JS I18N frameworks.

Presenters:

Shawn Xu
Internationalization Program Manager, Netflix

Tim Brandall
International Product Experience Manager, Netflix

Track 3 - **Pseudo Localization at Netflix**

Internationalization at Netflix starts with our UI design teams, before a single line of code is ever written. Hear about the tools and support the Netflix internationalization team provides to make sure new product UIs are world-ready. From pseudo localization plugins for common UI design tools, to on-the-fly machine translation for world proofing new UI concepts, we'll show you how it's done at Netflix.

12:00-13:00 - LUNCH

13:00-13:50	SESSION 10

Presenters:

Tex Texin
Chief Globalization Architect, Xencraft

Mike McKenna
Globalization Strategist, PayPal

Track 1 - Is Your Global Business at Risk?

There are many requirements to create a high quality global application that customers value. Failing to meet these requirements may hurt sales. More egregious failures may damage your brand. Some mistakes can result in legal action, penalties and even expulsion from a market.

Two of the industry's seasoned experts will describe globalization requirements, key performance indicators, and potential business risk. If you are unsure of the risks that errors in globalization can cause to your organization and how to avoid them, attend this session. This presentation will include, among other topics:

- Assessing the risk that your product isn't ready for global distribution
- Capturing metrics that quantify quality and assess whether your organization is improving
- Automating the collection of key performance indicators via automated systems where possible

---

Presenter:

Addison Phillips
Principal Globalization Architect, Amazon

Track 2 – A Man Called Horse: Working with User Generated Unicode Content

Unicode character is U+10085 LINEAR B IDEOGRAM B105M STALLION, or, in context, “horse”.

Presenter:

Moriel Schottlender
Senior Software Engineer, The Wikimedia Foundation

Track 3 - How We Let Our Users Translate Wikipedia's Interfaces for more than 400 Languages and Locales

People around the world read and edit Wikipedia in more than 400 languages and locales, using multi-lingual interfaces to interact with different content languages and produce articles that are read by billions all over the world. In order to foster the widest collaboration and contribution of the sum of all knowledge, our users can read a Wikipedia article in one language but view the interface in another, allowing them to read Wikipedia articles in English with a Spanish interface, or, even more complexly, an Arabic one.

Many languages have various properties that make translations complicated, like the use of gendered verbs, differentiated suffixes for numerical counts, and different rules for usage of singular versus plural words; these challenges require the system to allow for a high level of flexibility in translation, while still maintaining clear focus on what the interface message is aimed at.

And yet, with such a high number of languages to translate, the best translators come from among our own users, who are not necessarily technically savvy enough to install development environments or manipulate or edit code. We give them access to set and review translations through an interface editing system, and allow them to specifically articulate the best correct term for interface actions, be they generic and widespread or minute and specific.

This session will concentrate on the infrastructure that allows this process to happen; how do we define messages that can be adjusted to languages that have gendered pronouns and verbs? How do we account for differences in numerals, plural rules, or language variants? How do we give clear context to the message that is translated, so the translators can utilize the subtleties of their language to maintain clarity? How do we allow non-technical users to contribute translations in a flexible and yet controlled manner, into our production software, and what tools do we offer to make this process easy for other organizations that wish to follow our lead?

13:50-14:40 SESSION 11

Presenters:

Trevor Cortez
Localization Engineering Manager, Apple Inc.

Vivian Robison
Localization Tools Engineer, Apple Inc.

Fredrik Stenshamn
Linguistics Team Manager, Apple Inc.

Track 1 - Internationalization and Localization Techniques in Xcode

Xcode is a powerful tool used to create amazing apps for iOS and macOS. Xcode also makes it easy to design and test your application in multiple languages. This session will give you a tour of Xcode's internationalization and localization features, including finding localizability issues, managing localized content for different file types, previewing and testing your applications in different languages, and techniques for testing and handling bidirectional text.

- Ensuring your app is world ready, including internationalization best practices in Swift/Objective-c
- Before adding new languages to your app:
  - How to find localizability issues with the static analyzer
  - How to preview your app in different pseudo-language mode
  - How to preview layout in RTL mode
- Rendering bidirectional text in your app

- Prepping your app for localization
- How to export localizable content for translation
- Handling pluralization and width variants with stringsdict
- Handling non-string resources
- How to import completed translations back into your project
- How to make your app's UI automatically adapt for other languages
- Testing your app in different languages with XCUITest

Presenters:

Craig Cornelius
Software Engineer, Google, Inc.

Daan van Esch
Technical Program Manager, Google, Inc.

Luke Swartz
Product Manager, Google, Inc.

Track 2 - Supporting 1000+ Languages? Language Technology at Scale

Millions of people around the world speak a native language that is not well-supported by information technology products. Technology companies must make hard trade-offs when deciding which languages to support at what levels, but at Google we've started a variety of projects aimed at enabling better support for a large number of languages at scale. We'll discuss a range of efforts from Search to Input to encoding, and invite discussion about challenges and opportunities for the industry to move forward, and how together we can help large and small languages alike thrive using technology.

Presenter:

Nova Patch
Product Director, Shutterstock

Track 3 - Punctuation Internationalization

Punctuation is commonly just a bullet point in internationalization materials, if mentioned at all. This session will explore the depths of global punctuation and how to support dynamically localized punctuation throughout your products to provide a first-class user experience in all locales. Topics will include:

- A world tour of punctuation!
- To localize, or not to localize?
- A matter of style guides
- Punctuation inside (and outside) translations “Leveraging the Unicode CLDR – Beyond the CLDR: opportunities for standardization” Review of available software libraries

14:50 – 15:10 - Afternoon Refreshments

15:10 - 16:00	SESSION 12

Presenter:

Behdad Esfahbod
Software Engineer, Quora, Inc.

Track 1 - Introduction to Unicode & i18n in Rust

The Rust Programming Language has native support for Unicode Characters’ Unicode Scalar Values, to be exact. The language provides fast and compact string type with low-level control over memory consumption, while providing a high-level API and enforcing memory and data safety at compile time. The Rust Standard Library covers the basic Unicode functionalities, and third-party libraries – called Crates – are responsible for

the rest. UNIC's Unicode and Internationalization Crates for Rust is a project to develop a collection of crates for Unicode and internationalization data and algorithm, and tools to build them, designed to have reusable modules and easy-to-use and efficient API.

In this talk we will cover the basics of Rust's API for characters and strings, and look under the hood of how they are implemented in the compiler and the standard library. Afterwards, we look at UNIC's design model, how it implements various features, and lessons learned from building sharable organic micro components. The talk is suitable for anyone new to Unicode, or Unicode experts who like to learn about how things are done in the Rust world.

Presenters:

Vivekananda Pani
Co-founder & CTO, Reverie Language Technologies

Bhupen Chauhan
Technical Lead, Research, Reverie Language Technologies Pvt. Ltd.

Track 2 - Transliteration for Indic Languages

With a 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:

Zibi Braniecki
Senior Platform Engineer, Mozilla

Track 3 - Fluent 1.0 - Next Generation Localization System from Mozilla

Localization systems have been largely stagnant over the last 20 years. The last major innovation - ICU MessageFormat - has been designed before Unicode 3.0, targeting C++ and Java environments. Several attempts have been made since then to fit the API into modern programming environments with mixed results.

Fluent is a modern localization system designed over last 7 years by Mozilla. It builds on top of MessageFormat, ICU and CLDR, bringing integration with modern ICU features, bidirectionality, user friendly file format and bindings into modern programming environments like JavaScript, DOM, React, Rust, Python

and others. The system comes with a full localization workflow cycle, command line tools and a CAT tool. With the release of 1.0 we are ready to offer the new system to the wider community and propose it for standardization.

16:10 - 17:00	CLOSING SESSION

Moderators:

Martin Dürst
Professor, Aoyama Gakuin University

Alolita Sharma
Principal Technologist, Amazon Web Services (AWS)

Lightning Talks

This is the fourth installment of the very successful Lightning Talks from IUC 39-IUC 41. This 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. Proposals for lightning talks must be sent to the moderators, Alolita Sharma and/or Martin Dürst, by September 3rd. 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.