CONFERENCE PROGRAM

Wednesday, October 14, 2020

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
Sr. Technical Product Mgr., Amazon

Mike McKenna
Director World Ready Engineering, 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

Presenter:

Track 3: Put ICU to Work

Steven Loomis
Senior Software Engineer

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.

10:30-12:00	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** *Sr. Technical Product Manager, Amazon* **Mike McKenna** *Director World Ready Engineering, 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: **Mihai Nita** *i18n Senior Software Engineer, Google, Inc.*	**Track 3: Introduction to Android's Internationalization** This tutorial gives an introduction to the Android's internationalization and localization features, including a hand-on tutorial for developing an internationalized Android app from scratch (localizability, formatting, bidi, etc.) It is technical, there will be some code, but non-programmers should be able to follow.

12:00 – 12:30 – Networking Break – Chat & Q&A

12:30-14:00	SESSION 3 TUTORIALS
Presenter: **Josh Hadley** *Senior Computer Scientist, Adobe*	**Track 1 – Font Construction with AFDKO** The Adobe Font Development Kit for OpenType (AFDKO) is an open source toolkit for creating and manipulating OpenType fonts. In this hands-on tutorial, you will explore the relationships among Unicode, OpenType and fonts by building a working OpenType font. You will leave the tutorial with a firm understanding of the difference between characters and glyphs, as well as how Unicode text input is transformed into a visual depiction with fonts. Some familiarity with using command-line (Terminal) tools will be helpful. To be ideally prepared, you should bring a Mac, Windows, or Linux computer with a web browser, Python 3.6 or later, and the latest AFDKO installed. See https://github.com/adobe-type-tools/afdko for instructions and details. A link for remaining tutorial materials will be provided at the conference.
Presenter: **Tex Texin** *Chief Globalization Architect, Xencraft*	**Track 2 - Web Internationalization** This tutorial, updated in 2020, 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 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: **Jim DeLaHunt** *Principal, Jim DeLaHunt and Associates*	**Track 3 – Email Addresses and Domain Names are NON-Latin! Now What?** Email addresses, and domain names, are no longer limited to ASCII Latin script. They can now be http://普遍接受-测试。世界 or مانيش@أشوكا.الهند or données@fußballplatz.technology. Software, frameworks, and workflows will need to change to accommodate. What are Internationalized Domain Names (IDN) and Email Address Internationalization (EAI)? What do you need to know? What do you do next? This tutorial brings you up to speed. It explains IDN and EAI. It shows you the implications. It connects you to sources of information. It helps you understand what this will mean for you. Suitable for software developers, QA, marketers, system administrators, and management.

14:30-15:30	SESSION 4 TUTORIALS
Presenter: **Steven Loomis** *Senior Software Engineer*	**Track 1 – Submitting Locales into CLDR** The Common Locale Data Repository (CLDR) project from Unicode provides language and region-specific locale data and structure for software internationalization. But how do new locales get added? This tutorial will take you through the entire process of adding new data to CLDR starting from step zero, including planning and

community aspects, XML data for seeding new locales, and use of the CLDR Survey Tool for inputting content.

Presenter:

Dr. Martin Dürst
Professor, Aoyama Gakuin University

Track 2 – Character Equivalences, Mappings, and Normalization

The multitude of characters available in Unicode means that there are many ways in which characters or strings can be equivalent, similar, or otherwise related. In this tutorial, you will learn about all these relationships, in order to be able to better work with Unicode data and programs handling Unicode data. The tutorial assumes that participants have a basic understanding of the scope and breadth of Unicode, possibly from attending tutorials earlier in the day.

Character relationships and similarities in Unicode range from linguistic and semantic similarities at one end to the same character being represented in different character encodings or Unicode encoding forms at the other end. In the middle, numerical and case equivalences, compatibility and canonical equivalences, graphic similarities, and many others can be found. This sometimes bewildering wealth of characters, equivalences, and relationships is due to the rich history of human writing as well as to the realities of character encoding policies and decisions.

The tutorial will give some guidance to help users navigate equivalences and differences for their use cases and applications. Each of these many equivalences or relationships can or should be ignored in some processing contexts, but may be crucial in others. Contexts may range from use as identifiers (e.g. user ids and passwords, with security consequences) to searching and sorting. For most of the equivalences, data is available in the Unicode Standard and its associated data files, or is provided by other standards such as IDNA and PRECIS. But the use of this data and the functions provided by various libraries requires understanding of the background of the equivalences.

When testing for equivalence of two strings, the general strategy is to map or normalize both strings to a form that eliminates accidental (in the given context) differences, and then compare the strings on a binary level. The tutorial will not only look at officially defined equivalences, but will also discuss variants that may be necessary in practice to cover specialized needs. We will also discuss the relationships between various classes of equivalences, necessary to avoid pitfalls when combining them, and the stability of the equivalences over time and under various operations such as string concatenation.

Presenters:

Denver Lin
Software Engineer, Citrix

Marshall Wu
Software Development Manager, Citrix

Track 3 – Introduction of Multilingual Text Input in Linux

This tutorial helps you understand how the text input of different languages are supported in Linux, includes Latin and non-Latin scripts.

The knowledge of X Keyboard system, Xkb Extension, as well as the keyboard layouts definition in Xkb will be included in this tutorial. You will learn the power of X Keyboard system, which supports almost all kinds of keyboards and can be easily extended or customized. This tutorial also addresses the XIM (the Input Method Protocol for X), which is the most important protocol being used for almost all non-Latin scripts input on Linux. One of the mostly adopted implementation of XIM, IBus, will be introduced in this tutorial.

Thursday, October 15, 2020

08:45-09:00	Conference Welcome & Opening Remarks
09:00-10:00	**SESSION 1**

Presenter:

Claudia Galván
Technical Advisor, Early Stage Innovation

Track 1 - Privacy Best Practices Around the World

The General Data Protection Regulation (GDPR) and the California Privacy Act (CCPA) have raised the bar on privacy in Europe, and new privacy laws are in effect worldwide. 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 in an extension of previous talks and will provide a high-level overview of managing privacy in global products.

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

Presenter:

Mike McKenna
Director World Ready Engineering, PayPal, Inc.

Track 2 – CLDR: What's in a Personal Name?

"It would be helpful if CLDR could provide some locale-specific information on the structuring and use of personal names."

The content and structure of personal names can vary widely from region to region and locale to locale. This session will cover the current standards that exist in LDAP, hcard, and HTML as well as various commercial implementations. Then a look at real word name examples, ranging from mononyms in Indonesia to patronymic and matronymic names in Iceland, Spain and Portugal to Arabic ancestral naming practices. We will look at name usage in different legal, business, familial and formality contexts. Finally, we will present the current state of the evolving effort to add locale-specific name structures to CLDR.

10:00-11:00	**SESSION 2**

Presenter:

Riitta Koikkalainen
Information Specialist, National Library of Finland

Track 1 - New Regulations on Accessibility – Searching for Best Practices in the Worlds of Localization

What does accessibility mean, and what does it not mean, in the worlds of internationalization and localization?

In our presentation, we explore some impacts of the new legislation and regulation on ICT, mainly user interfaces and user experiences. In Europe, the themes of accessibility and ICT have become once again subjects of importance in debates since 2016 due to two directives given by European Parliament and Council:

- Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies

- Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services

Intention behind these regulations is, in the end, to increase equality. The aim, originally, has been to guarantee similar user experience on the web for each group, for example people with visual challenges. As a bonus, all users benefit: the ideas of the directives, when brought into practice, ease the usage of ICT, including electronic equipment, for each and every user.

In the title, we mention the search for best practices. We must all make better, that is, accessible, ICT services and products. The goal is mutual, we are in this together. To reinforce our point, there will be a small demonstration at the end of our presentation.

Presenters:

Zibi Braniecki
Senior Staff Platform Engineer, Mozilla

Shane Carr
Senior Software Engineer, Internationalization, Google, Inc.

Nebojša Ćirić
Software Engineer, Google, Inc.

Track 2 – ICU4X: Solving i18n on Client-Side Platforms

What does accessibility mean, and what does it not mean, in the worlds of internationalization and localization?

In our presentation, we explore some impacts of the new legislation and regulation on ICT, mainly user interfaces and user experiences. In Europe, the themes of accessibility and ICT have become once again subjects of importance in debates since 2016 due to two directives given by European Parliament and Council:

- Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies
- Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services

Intention behind these regulations is, in the end, to increase equality. The aim, originally, has been to guarantee similar user experience on the web for each group, for example people with visual challenges. As a bonus, all users benefit: the ideas of the directives, when brought into practice, ease the usage of ICT, including electronic equipment, for each and every user.

In the title, we mention the search for best practices. We must all make better, that is, accessible, ICT services and products. The goal is mutual, we are in this together. To reinforce our point, there will be a small demonstration at the end of our presentation.

11:00-12:00	**SESSION 3**

Presenter:

Makara Sok
Project Support Specialist, SIL International

Track 1 – Discrepancies in Khmer Unicode Character Ordering Rules and a Proposed Solution

Unicode Standard, OpenType and Open Forum of Cambodia have been implementing different character orderings for Khmer script. This discrepancy causes various issues, such as: confusability, vulnerability, searchability and compatibility. To resolve these issues, a set of new character ordering rules is proposed and some recommendations have been made so that the rules are inclusive and complete. The proposal is based on a thorough analysis of character usage in all contemporary and some historical languages using the script.

Presenters:

Sreejita Dutta
Software Engineer,
PayPal Inc.

Ritika Mathur
Software Engineer,
PayPal Inc.

Mike McKenna
Director World Ready Engineering,
PayPal, Inc.

Track 2 – Address Details Are Not So Simple

The customer's address plays a crucial role in payment transactions or placing orders. Having a reliable address not only provides proof of the customer's identity but also helps reduces fraud and money laundering. Hence, it is crucial for a customer's address to reflect reality. However, when you are required to gather detailed address information and make it intuitive and familiar to the user anywhere in the world, the problem gets a lot harder.

We thought we understood, based on available standards and repositories, how every region should have its layout formatted, and what was and was not required. We were wrong. We will be discussing along with lots of examples:

- Detailed address formats and how they vary from region to region
- Our experience with available international and regional standards, and how we vetted our designs to real-world usage
- Input layouts, display layouts, - and validations
- Flexible data structure design to accommodate the variability from region to region
- Transform design to be sure data is not lost when moving between detailed and less detailed schemas and how to extract relevant information for compliance needs
- Support for multi-lingual countries where layouts may change depending on the script used or the prevailing language of the region
- Generating and using Test data

12:00-13:00	**SESSION 4**

Presenters:

Sreejita Dutta
Software Engineer,
PayPal Inc.

Ritika Mathur
Software Engineer,
PayPal Inc.

Lucas Welti
Globalization Architect,
World Ready Engineering,
PayPal, Inc.

Track 1 - Is Your Application World Ready?

As we build new applications, we need to ensure that they are developed in a way that can be easily expanded to new regions and languages. The earlier we can detect poor development practices, the better. For that reason, we have built two major plugins to use in development and continuous integration processes.

For JavaScript issues, we have created an Eslint i18n Plugin to detect bad practices in the Code. Using our Eslint plugin, we are able to detect inline content, hard-coded date formats, missing parameters, use of third party i18n Libraries and third party L10n Libraries. For source content, we have built a L12y plugin to detect bad practices in messages intended to be localized. With this plugin, we can detect partial or fragmented sentences, embedded content that should be formatted differently depending on the locale such as phone numbers, dates, and URLs, and other issues.

In this session, we will discuss the prevalent problems we encountered that precipitated the creation of these plugins, issues with false negatives, and integration into the development process to enable feedback to the individual developer as well as centralized monitoring so we can track and alert product teams that need help.

Presenters:

Zibi Braniecki
Sr. Staff Platform Engineer, Mozilla

Mark Davis
Internationalization Architect, Google, Inc.

Steven Loomis
Senior Software Engineer

Track 2 - CLDR Panel: CLDR Key Users and Contributors Answer Your Questions

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

13:00 – 13:30 – Networking Break – Chat & Q&A

13:30-14:30 **SESSION 5**

Presenter:

Jim DeLaHunt
Principal, Jim DeLaHunt and Associates

Track 1 - Internationalizing Date-time API Consistent with the Earth, Moon, and Leap Seconds

When does a minute have 61 seconds? Why does Posix time conflict with UTC? Why will California's day be 25 hours long on November 1 2020? And why will that same day be shorter in the Yukon? If we don't know how many days will be in this month, is that a problem? If we know our clocks could show the sun rising at 12:00 noon, many years in the future, should we avert that now? Will your software stumble at the next leap second? Should we abolish leap seconds altogether? Internationalization is about making software constructs match human requirements. Text strings and fonts get the glory, but time and date structures also need localization, and in surprising ways.

This presentation digs into the role played by Earth and Moon orbits, and human choices, in the various data structures returned by your platform's time() API. Suitable for all attendees (with plenty of links to specific specs and standards to satisfy the developers).

Presenters:

Zibi Braniecki
Sr. Staff Platform Engineer, Mozilla

Steven Loomis
Senior Software Engineer

Robert Melo
SW Internationalization Engineer, Motorola Mobility

Track 2 – ICU Panel: ICU Expert Implementers Answer Your Questions

ICU (the International Components for Unicode) is a widely-used implementation of Unicode. ICU will be introduced briefly, and then we continue with a panel discussion focused on the experience of direct consumers of (and contributors to) the project.

Topics discussed will include benefits and challenges of using ICU.

This discussion, moderated by Steven Loomis and Markus Scherer, will allow plenty of time for questions from the floor, and general Q&A.

14:30-15:30	**SESSION 6**

Presenter:

Ben Yang
Director of Technology, PanLex The Long Now Foundation

Track 1 - Quick, How Many Characters am I Holding Up? Why Determining the Length of a String is Way More Complicated Than You Think

There are many ways of determining the "length" of a string, but most of them don't correlate well with an average person's expectations. For example, depending on encoding and what's being measured, the length of the Vietnamese character "å" could be 1, 2, 3, 4, 5, 6, 8, 12 or 16! But with the use of Unicode's "Extended Grapheme Cluster" algorithm, the length matches user expectation, and is reliably 1. This session will go over the following possible "lengths" of a string, with their pros and cons:

Bytes, with info on the differences between UTF-8, UTF-16, and UTF-32
Code Points, and why they don't always correspond to a "character"
(Extended) Grapheme Clusters, how their boundaries are determined, and how to implement them in a variety of programming languages
Considerations for cases beyond Grapheme Clusters

Presenters:

Macie Chervunkong
Director, Nyiakeng Puachue Hmong Committee

Dr. Craig Cornelius
Sr. Software Engineer, Google, Inc.

David Lee
Technical Lead, Nyiakeng Puachue Hmong Committee

Track 2 - Panel: Leveraging Digitization to Accelerate Language Sustainability

According to Wikipedia, there are between 6,000 to 7,000 languages currently spoken and up to 50% of them will have become extinct by the year 2100. These endangered languages are moving towards becoming extinct as speakers of the languages die with the new generation no longer speaking their own languages and shifting to other larger languages related to mass economic changes of globalization, political pressures and lack of resources to learn from.

This panel discussion, moderated by Alolita Sharma, Director, Unicode Consortium, will focus on leveraging digitization efforts on the Internet to accelerate language sustainability. The panel will also discuss revitalization challenges, and solutions. Solutions can include dissemination of know-how to set up language communities around endangered languages with language sustainability kits, leveraging Unicode projects to help preserve language encoding, developing fonts and keyboards for language scripts, extending full NLP support beyond the top 5% of languages to enable language usage with features like autocorrect, voice recognition and developing digital content to ensure endangered languages have a chance to capture their finer details for the next generation to learn and use. Come join in for a great conversation with experts using technology for language sustainability.

15:30-16:30	Birds of a Feather / Q&A / Reception
16:30-17:30	Wrap-up & Q&A

Friday, October 16, 2020

08:45-09:00	Opening Remarks
09:00-10:00	**SESSION 7**

Presenter:

Joel Sahleen
Globalization Architect, Domo, Inc.

Track 1 - International Languages and Regional Locales

This session looks at how SaaS development teams can enable international growth and improve localization fit, without increasing translation cost or compromising developer productivity. More specifically, it explains how development team can do this by using freely available internationalization libraries and localization data to expand their localization offerings vertically, from international languages to regional locales. The session is based our experiences with vertical locale expansion at Domo, where we are using ICU4J, CLDR, the JavaScript Intl API, moment.js and a custom-built internationalization framework and webpack loader to localize our web user interface into two dozen regional locales that are generated from six parallel sets of international language files. The presentation is divided into four parts. It will start with an overview of the globalization problems we are trying to solve at Domo and a review of some analytics that lead us to believe vertical locale expansion could help us extend our market reach and better serve our international customers. This will be followed by a discussion of the different ways vertical locale expansion can be implemented and the things we took into consideration when designing our locale-oriented internationalization and localization framework. Next, we will turn to the implementation of this framework itself and go through some of the lessons we learned during its construction. The presentation will conclude by briefly summarizing the advantages and disadvantages of vertical locale expansion as a means of enabling international growth and improving localization fit.

Presenters:

Shane Carr
Senior Software Engineer, Internationalization, Google, Inc.

Ujjwal Sharma
Compilers Hacker, Igalia

Track 2 – What's New in ECMAScript 2020 Intl

JavaScript continues to dominate the programming languages popularity rankings. The language that powers the client-side web is also increasingly popular on the server side and is now used to develop on mobile devices, IoT, cloud services, and many others.

ECMAScript is the standard powering the language governed by the Technical Committee 39 (TC39). TC39-TG2 is a subgroup of that committee standardizing ECMA-402, which focuses on bringing Internationalization capabilities to JavaScript.

In this presentation we will guide the audience through the recent additions to the language such as Intl.Locale, Intl.RelativeTimeFormat, Unified Intl.NumberFormat, and a number of upcoming ones such as Temporal calendars, Intl.ListFormat, Intl.Segmenter, Intl.DurationFormat and Intl.DisplayNames.

Presenter:

Daniel Yacob

Director, Ge'ez Frontier Foundation

Track 3 - EMUFI and the Ethiopic Endgame

The digitization of Ethiopic manuscripts has exploded over the last two decades since the Ethiopic script joined the Unicode Standard. Surprisingly, these efforts to preserve the manuscript heritage have been largely limited to image scanning. Leaving the content inaccessible to text search and analytics. A major barrier to text representation is the absence from standards of over 300 symbols for letter forms, numeral variants, punctuation and musical notation found in the ancient and mediaeval manuscripts corpus. Closing this gap is now the Ethiopic script endgame. Enter EMUFI.

The Ethiopic Manuscript Unicode Font Initiative (EMUFI) brings manuscript scholars and font designers together to tackle the remaining gaps to fully digitizing Ethiopic manuscripts as textual documents. The EMUFI project draws inspiration from, and indeed is modeled after, the successful "Mediaeval Unicode Font Initiative" (MUFI) for Latin script. The effort aims to review unencoded symbols for Ethiopic and arrive at cooperative ways of supporting them in font technology such as under stylistic alternatives, private use ligatures, or as entirely new symbols in need of standards encoding.

The presentation will review the EMUFI effort, the institutes involved, as well as review the character collections categorically along with their relevance to literature preservation in the digital age.

10:00-11:00 SESSION 8

Presenter:

Nova Patch
Director, Internationalization & Localization, Shutterstock

Track 1 - 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

Presenters:

Behnam Esfahbod
Software Engineer, Quora, Inc.

Track 2 - Composable i18n API in JSX/React

React is a widely adopted rendering framework for web-based applications. One of the main reasons for that is the architecture for developing reusable functional-style components as building blocks. Most frameworks available to enable international support for such applications, however, miss the reusability aspect of this

Chris Zeng
Software Engineer,
Quora, Inc.

model, enforcing products to be minimal and over-simplified in textual communications with their users. In this talk we introduce an internationalization framework for React applications, enabling reuse of messages (templates) via composition. The JSX-based syntax relieves developers from learning a new template syntax for marking up interface messages, while the static code analyzer ensures the templates are identified and processed; and finally, the runtime libraries handle translation delivery and rendering in the target language.

Quora is a platform that empowers people to share and grow the world's knowledge, and is available in 24 languages and 11 writing systems.

Presenter:

Neil Patel
Director, Partner,
JamraPatel

Track 3 - The Modern Writing Systems of West Africa: Support After Unicode

Recent vigorous script encoding activities have brought many new writing systems into Unicode. Amongst these are African writing systems that have only been in existence for less than a century. The relative youth of these scripts, combined with the language and cultural dynamics in Africa, brings unique challenges when it comes to supporting them in technology. The common conception is that technology inherently aides in the spread of underserved writing systems. While this is true, entrenched paradigms within the tech community can, in fact, impede progress.

We will look at what happens after a script is encoded, based on our experience in working with language communities to make and distribute apps that support commerce, basic communication and native content creation. Focusing on N'ko and Adlam, two of West Africa's most active new writing systems, we will review the current state of their support in computing, dive into the complex landscape in which these writing systems exist, and explore the difficulties in the road ahead to achieving broader support.

11:00-12:00 SESSION 9

Presenters:

Craig Cummings
Senior Technical Product Manager, Amazon

Mike McKenna
Director World Ready Engineering, PayPal, Inc.

Tex Texin
Chief Globalization Architect, Xencraft

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. 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

- Performance tradeoffs including: client vs. server-side processing, footprint vs. speed, searching, sorting, and considerations for WAN networks, low performance networks, and CDNs
- Installation considerations including: individual vs. multiple language vs on-demand installations, update processes, practices with stores (Google, Apple, etc.)

Presenter:

Jennifer Daniel
Unicode Emoji-Subcommittee Chair, Google, Inc.

Track 2 - Race is Not a Skin Tone, Gender is Not a Haircut, Designing Emoji at Scale

Sometimes at engineering driven companies there can be a preconceived notion there is a "right" or "wrong" way of designing. This talk will explore how the emoji program operates in the spectrum between this false binary. After all, if race is not a skin color and gender is not a haircut how do you communicate the "idea" of "black woman" at emoji sizes? We'll discuss things like how skin tone was added to the standard, how they are working, and what the future holds.

Presenters:

Dr. Craig Cornelius
Sr. Software Engineer, Google, Inc.

Dr. Stephen Morey
Associate Professor, LaTrobe University

Track 3 - Connecting the Dots with Unicode - Bringing Unicode to Life in Assam

Although Unicode standard scripts have been defined for all of the Tai languages of Assam for around a decade, the actual adoption of Unicode for these Tai languages has been very slow.

From the perspective of script, there are two main groups of Tai people in Northeast India. (1) The Tai Khamti, Phake, Aiton and Khamyang all use a script which they call To Lik Tai (Tai letters), a script based on Shan and which is encoded as part of the Unicode Myanmar block. These languages are all still spoken, although Khamyang is very much endangered and very few members of the Aiton community are learning the script. (2) The Tai Ahom, who no longer speak their language as a mother tongue, and whose script has its own block within Unicode.

The challenges for the users of both of these scripts include i) Enabling these fonts on laptops and mobile devices, and across multiple applications and programs. ii) Producing and enabling keyboards that can access these fonts. While considerable progress has been made on both of these fronts, the implementation of these Unicode across the board is still not complete. For example, among the users of To Lik Tai (Khamti, Phake, Aiton and Khamyang), there has long been a preference for a printed form that includes "black dots" as part of the shape of consonants and some vowels. Characters of this kind are supported in Unicode as "variants" of the basic forms. In February 2020, community members responded to text on a social media site that did not display the dotted forms, indicating that "this isn't our script," Users are waiting for fonts and keyboards that support the variants and input methods that let then create text with the right codes.

This talk will present some approaches in progress, and also will discuss some of the issues and challenges faced by the community, by a linguist, and by technology providers in bringing Unicode to life (or "Unicode to life") in Northeast India.

12:00-13:00	SESSION 10

Track 1 - Cancelled

Presenter:

Marek Jeziorek

Technical Program Manager, Google LLC

Track 2 – Noto Fonts: Striving for Excellence and Coverage

The Noto project attempts to develop a unified set of typefaces for everything defined in Unicode -- all open source. Noto fonts are in daily use by billions of people because they ship in every Android and ChromeOS device, and many designs have versions that are adjusted to work within the vertical space constraints of labels and buttons. Now with coverage up to and including Unicode 12+ (for some scripts), the set is intended to be visually harmonized. This presentation provides insights into never ending strife to cover new and emerging scripts while at the same time improving the fonts from quality point of view. We will also describe the Noto font creation process, from the design brief with letter-form characteristics to how the fonts are built and tested.

Presenters:

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

Charles Riley
Catalog Librarian, Yale University Library

Dr. Deborah Anderson
Researcher, Dept. of Linguistics, UC Berkeley

Track 3 - Script Encoding and Beyond in 2020: Challenges and Successes

This talk provides an overview of the efforts by the Script Encoding Initiative (SEI) to expand support for the world's writing systems in Unicode in 2020. It will also present an in-depth report on script-encoding proposals that have posed unique challenges, namely Old Uyghur and Proto-Elamite, the difficulties in encoding newly-created scripts, such as Tangsa, and the hurdles for encoding other scripts, specifically those in West Africa.

Work on scripts in 2020 has included proposals that request additions to already encoded scripts, such as Arabic additions for the Quran or Latin extensions for IPA, as well as documents that propose adjustments or improvements to encoded scripts. An example of the latter is ongoing work to improve the Mongolian script model. In general, authors requesting additions have been able to take advantage of the wealth of successful proposals to use as a guide, as well as getting input from the Script Ad Hoc, a group of experts that meets monthly and provides feedback to proposal authors and makes recommendations to the Unicode Technical Committee.

New script proposals are being submitted at a relatively slower rate than in the past, in part because the remaining unencoded scripts require more research and their features present complex challenges. Although consisting on 18 letters, the Old Uyghur script was adapted in different ways across different Central Asian communities over time, resulting in changes to the repertoire, letterforms, and creating of new signs. Encoding Old Uyghur has required rethinking the model used for cursive-joining scripts in the standard. On the other hand, Proto-Elamite is a partially deciphered script, and the language it was used to represent is still unclear. Still, Proto-Elamite adopted most of the numerical signs and numerical systems from Proto-Cuneiform, which provides insight into some signs. Issues relating to encoding this script involve questions on the script's relation to Proto-Cuneiform and the already encoded Cuneiform script, as well as making a case for a script that is not fully deciphered. For relatively new scripts, such as Tangsa and Beria, evidence is needed showing that the script is widely accepted by the intended user community, that it is in use, and that the writing system is stable.

Another focus of work on scripts by SEI and other groups, such as TranslationCommons and Athinkra, has

been to try to help ease the transition for communities from Unicode code points to Unicode-enabled keyboards and fonts on computers and devices. TranslationCommons is producing a guide for language communities on the steps involved. Work on fonts by Noto, ANRT, and other entities is helping to streamline font creation. However, additional components are also needed, including testing implementations of scripts – especially complex scripts – and submission of locale data for modern languages. Athinkra, under Charles Riley's guidance, has been working to get West African scripts supported on devices.

This presentation on the important task of expanding the Unicode standard to support more and more of the world's writing systems will appeal to both experts of Unicode and those new to the world of character encoding.

13:00 – 13:30 – Networking Break – Chat & Q&A

13:50-14:40	SESSION 11
Presenters: **Zibi Braniecki** *Sr. Staff Platform Engineer, Mozilla* **Elango Cheran** *Software Engineer, Google, Inc.* **Mihai Nita** *I18n Senior Software Engineer, Google Inc.*	**Track 1 - MessageFormat 2.0: Standardizing Localized Message Strings Across the Industry** MessageFormat Working Group has been tasked with developing a successor to ICU MessageFormat based on experience from projects such as Fluent, FBT, Siri, I18Next, and others. In this presentation, the audience will be guided through the design principles, core goals and the roadmap of the project. We'll also look for early feedback and will provide an introduction on how to join the effort. ICU MessageFormat has been an important part of the internationalization ecosystem for over 20 years now. It is used not only for text formatting but also in some localization contexts. Over the last several years, there has been a renewed energy around software localization coming from an increasingly multilingual Web. There has also been a growing demand for high-quality complex translations coming from new UX such as voice assistants, social networks, and interactive UIs. This new generation of use cases has led to a high level of interest in how MessageFormat can be designed to support these diverse needs.
Presenters: **Cosimo Lupo** *Software Engineer, Google, Inc.* **Dominik Röttsches** *Software Engineer, Google, Inc.* **Roderick Sheeter** *Staff Software Engineer, Google Fonts TL/M, Google, Inc.*	**Track 2 - Color Vector Variable Fonts** OpenType (main active font format) support for color vector fonts is weak. Multiple vendors, including Google, have long been limited to delivering a bag of bitmaps wrapped up as a font. Emoji webfonts have also been a challenge as the long and overlapping RGI sequences do not lend themselves to delivery with unicode-range, plus bags of bitmaps lead to very large files. We will discuss progress toward extending OpenType and the Open Text Stack to support color vector variable fonts, application to Emoji, and a path to efficient incremental web delivery (general purpose, but with a focus on Emoji).

Presenters:

Dr. Craig Cornelius
Sr. Software Engineer, Google, Inc.

Muhammad Noor
Co-Founder and Managing Director, Rohingya Project

Track 3 - Rohingya and Unicode: How Language Technology Helps a Community Living in Exile

Rohingya Language has been spoken and written for centuries in the Rohang Valley of today's Rakhine state of Myanmar, formerly known as Arakan. Rohingya language engravings have been found in the hills of Mrauk U, the ancient capital of the Arakan kingdom. Throughout the centuries, Rohingya writing has taken many forms and shapes due to the influence of kingdoms and emperors. In 1980, a modern script was introduced by a group of scholars led by Molana Mohammed Hanif. The new script includes features of some ancient symbols and tone. Since its development, books, history, culture, news and many other contents have been written in this script. Today hundreds of thousands can read and write in Rohingya language throughout the world. In 2000 the first computer font was developed by Muhammad Noor to bring digital presence, and the Rohingya script was added to Unicode in 2017.

The authors describe the history of Rohingya writing, the modern script and its usage, the standardization efforts, and recent technology support for education, communications, culture, and social action for the Rohingya people.

14:30 - 15:30 SESSION 12

Presenters:

Martin Reddy
Engineering Manager, Apple, Inc.

George Rhoten
Language Technologies Developer, Apple, Inc.

Track 1 - Authoring Grammatically Correct Conversational Templates for Siri

Formatting messages to display in a user interface has improved in capability over the years. For example, CLDR and ICU have provided the ability to format dates and numbers for various languages and regions. They have provided the capability to choose the correct singular, dual and plural forms of phrases for various languages. They have provided the capability to choose the correct grammatical gender or natural gender of phrases for various languages. Much of this functionality works well when the set of words used in the message are known when translating into another language. Unfortunately, the available message formatting frameworks fail to create grammatically correct sentences when messages reference user defined vocabulary or must be spoken in addition to being printed. This presentation covers the functionality in the message formatting framework of Siri that was developed to overcome these challenges. For example, pronouncing the number 1 in various contexts is a lot harder than writing it. Adding articles and prepositions is hard when the noun being referenced in a message is only known after the message has been translated. Some messages may need the plural form of user vocabulary that is originally provided in the singular form. Examples will be provided that illustrate how phrases naturally used by humans remain difficult to handle in traditional message formatting frameworks.

Presenter:

Elango Cheran
Software Engineer, Google, Inc.

Track 2 - Deriving Lexical Data for Tamil from Scratch Using Morphology

Lexical data for a language is often necessary to enable more advanced NLP work. New datasets are being created, but often they may not have sufficient POS-tagging, a permissive enough open-source license, or both. This talk offers a case study of how these issues influenced a project to derive a basic lexical dataset for Tamil by starting only with a word list, which was taken from Unicode CLDR's Unilex project's repository.

While ICU successfully handles a range of useful operations across the set of locales in CLDR, we already require lexical data in some cases for word breaking. Beyond that, for most languages, operations like spell-

checking, grammar checking, or even stemming require some form of lexical data. In turn, they may be initial requirements to more higher-level applications like Machine Learning algorithms.

The Unilex data input, Tamil lexical data inference code, and data output are made available under the Unicode license, which supports a wide variety of use cases with minimal restrictions.

This talk is intended for a general audience and provides a gentle introduction to the basic linguistic concepts related to the challenges and solutions. The fact that Tamil is both an agglutinative language and written in an abugida script poses specific challenges, and the code handles them through encoding of morphology rules and decomposing words into phonemes. The lessons learned through this project will hopefully provide insight for future work, especially for languages with similar writing systems and/or similar grammatical traits.

Presenter:

Liang Hai
Independent Researcher, (in cooperation with Typotheque)

Track 3 - An Open Knowledge Base for Indic Text Shaping

Introducing an ongoing cooperation between the speaker and Typotheque / Peter Bil'ak: "Indic text shaping for type designers" (https://github.com/typotheque/text-shaping), a freely available and community-driven knowledge base. The project aims at providing the significant missing link of information for Indic scripts, between the low-level Unicode Standard and the incomplete OpenType Layout specifications.

Currently, only a small group of developers hold the critical information of how Unicode Indic text can be correctly transformed into intended shapes, while the whole industry and numerous user communities heavily rely on them. With an insightful analysis model establishing the theoretical framework, this initiative is able to coherently offer valuable data and practical tutorials. Thus everyone from average native users to foreign type design professionals is enabled to quickly produce a working Indic font, as how it should have been.

The speaker also hopes the expert community can gradually refine the presented analysis model, and reach a wide agreement on how to analyze Indic text's encoding and shaping. This will allow us to contribute back to and improve the Unicode Standard.

15:30 - 16:30	**Lightning Talks**

Moderators:

Martin Dürst
Professor, Aoyama Gakuin University

Alolita Sharma
Director, Unicode Consortium

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.

16:30-17:30	Wrap-up & Q&A