Internationalization & Unicode® Conference

OCTOBER 13-15, 2021
SANTA CLARA, CA

CONFERENCE PROGRAM

Wednesday, October 13, 2021

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
Director World Ready Engineering, PayPal, Inc.

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
- International Components for Unicode (ICU)
- Unicode Implementation Details and Recommendations
- The Unicode Consortium umbrella of Unicode, that is CLDR, ICU, and more
- 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
- Unicode implementations on practically every modern device - in operating systems, browsers, applications, programming languages, and more
- How Unicode is Evolving

Presenters:

Steven Loomis
Senior Software Engineer

Craig Cornelius
Senior Software Engineer, Google, Inc.

Track 3: The ICU Tutorial

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 to include: Installation (C++ libraries, Java .jar files, Java SPI for JDK integration), verification of installation, introduction and detailed usage analysis of ICU's frameworks (normalization, formatting with the fluent API, calendars, collation, break iteration, Unicode properties, transliteration). We will also cover the packaging of ICU data, integrating ICU into an applications development process, and how to get involved in the ICU development community.

10:30-11:00 - Morning Refreshments

11:00-12:30 SESSION 2 TUTORIALS

Presenter:

Track 1: An Introduction to Writing Systems & Unicode Part 2

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: Unicode in Action

Craig Cummings
Senior Product Manager, Amazon

Mike McKenna
Director World Ready Engineering, PayPal, Inc.

Tex Texin
Chief Globalization Architect, Xencraft

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:

Craig Cornelius
Senior Software Engineer, Google, Inc.

Mark Durdin
Keyman Team Lead, SIL International

Joshua Horton
Keyman Predictive Text Lead, SIL International

Steven Loomis
Senior Software Engineer

Track 3: CLDR Keyboard Initiative and Implementation

The CLDR Keyboard project is an initiative to collect layouts and transforms of user input tools in a standard format. This tutorial will first describe the basic structural components of a keyboard definition as specified in UTS #35 LDML (https://unicode.org/reports/tr35/tr35-keyboards.html). The tutorial will demonstrate a basic layout for a single layer keyboard using existing tools such as visual editors. This will be extended to include shift, control, and other layers. The example will then be enhanced with transform rules to illustrate code substitution and code point reordering. The tutorial will also review some methods to implement such CLDR keyboards on digital platforms.

Additional details of the LDML capabilities will be reviewed to implement additional features needed in keyboard implementations. Attendees will also learn about tooling for creating platform-specific keyboard applications or modules that can be installed on user devices. Tools such as KeyMan Developer can be used to prototype implementations before exporting to formats for the CLDR Keyboard database. Finally, the tutorial will outline the process of proposing and adding new items to the CLDR keyboard repository.

12:30-13:30 - LUNCH

13:30-15:00 SESSION 3 TUTORIALS

Presenter:

Addison Phillips
Senior Principal Engineer, Amazon.com

Track 1 - Internationalization: An Introduction Part 1

Your global customers live in a culture you don't understand. How can you possibly build them a software solution--from the tiniest app to the largest enterprise system--that meets their diverse needs, if you can't even speak their language?

Internationalization enables software teams, large and small, to ship software that delights a global audience. This tutorial presents the basic concepts, with a focus on real world examples, so you can understand how to analyze a product for internationalization issues, develop a design or approach, and deliver a global-ready solution. You'll learn how to manage character encodings, deal with formatting and presentation, untangle bidirectional text, and get your product localized into many languages.

Presenters:

Craig Cummings
Sr. Product Manager, Amazon

Mike McKenna
Director World Ready Engineering, PayPal, Inc.

Tex Texin
Chief Globalization Architect,

Track 2 – Software Architect Considerations in Global Application Design

Presented by software internationalization experts, this session describes the typical architectural tradeoffs and considerations for 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,

Xencraft

public and commercial libraries, and modular design.
- 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 and publishing considerations including individual vs. multiple language vs on-demand installations, update processes, practices with app stores (Google, Apple, etc.)

Presenter:

Joshua Hadley
Senior Computer Scientist, Adobe, Inc.

Track 3 – 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, knowledge of advanced typographic features, and insights into how Unicode text input is transformed into a visual depiction using font data.
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 additional tutorial materials will be provided at the conference session.

15:00-15:30 - Afternoon Refreshments

15:30-17:00 **SESSION 4 TUTORIALS**

Presenter:

Martin Dürst
Professor, Aoyama Gakuin University

Track 1 – 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.

Presenter:

Mihai Nita
Senior Software Engineer, Google, Inc.

Track 2 – Android internationalization: An Introduction

This tutorial gives an introduction to the Android's internationalization and localization features, basic level. It includes a hand-on coding session, developing a well 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.

Presenter:

Ben Yang
Linguist/Software Engineer, Adobe, Inc.

Track 3 – This Way, That Way, and Topsy-Turvy: Next Direction in Web Design

To create a truly multilingual web page or webapp, you’ll need to be prepared for text that doesn't just go left-to-right, but right-to-left, or even top-to-bottom! Thankfully, modern web standards are here to help, and it's easier now than ever to support languages that are read and written in directions other than that of the Latin alphabet. This tutorial will explain how to use modern HTML and CSS to support RTL and TTB text.

Some topics that this tutorial will cover include:

- An overview of the diversity of writing directions
- Basic CSS properties for writing directions
- Logical vs. physical properties
- Using flexbox and CSS grid in multidirectional environments
- Browser support of multidirectional text properties
- Where browsers fail with multidirectional text, and how they can be improved

Participants are expected to have some basic familiarity with HTML and CSS.

Thursday, October 14, 2021

08:45-09:15 ***WELCOME & OPENING REMARKS***

09:15-10:00 **KEYNOTE PRESENTATION – Taking Playfulness Seriously – When Character Sets Are Used in Unexpected Ways**

Presenter:

Gretchen McCulloch
Internet Linguist

When encoding characters, it's easy to look at formal reference documents: dictionaries, descriptive grammars, and other published materials. But most of the writing on the internet doesn't go through editorial standardization, which means that users can pick up characters originally encoded for one purpose and repurpose them for something else. By their very nature, there isn't a finite list of Unicode subversions, but some examples include: ASCII art, zalgo, glitch text, kaomoji, math mode alphabets used as ad-hoc fonts, and emoji sequences used as graphic elements.

Problems then arise when creative sequences are parsed by tools which have only been designed for conventional sequences, for example when ASCII art is read by screenreaders, kaomoji get broken across linebreaks, or emoji need to be rendered in LaTeX. Current workarounds are chiefly to screencap creative sequences and replace them as images with alt text, which defeats the many benefits of using Unicode characters in the first place (interoperability, searchability, editability, copy-paste-ability, etc.). These workarounds are a reasonable stopgap when a creative sequence is new and it is unclear whether it will persist; however, given that several are now up to decades old, I explore what a more systematic approach to the category of playful sequences could look like.

10:00-10:30 - Morning Refreshments

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

Presenter:

Martin Dürst
Professor, Aoyama Gakuin University

Track 1 – Slicing and Dicing Unicode Properties

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

This presentation introduces a library in the form of a DSL (domain-specific language) for handling Unicode property data. The language allows to derive or indicate property metadata (e.g. whether a property has binary or numeric values), to associate properties with data files, and to select subsets of properties or property values. The library is written in the programming language Ruby, and therefore easily allows to tie in additional processing steps at various stages.

We will show various applications of this DSL, from quick checks to the creation of compact data structures. We will also present ideas of how metadata about properties might be published in a more streamlined way.

Presenter:

Joel Sahleen
Globalization Architect, Domo, Inc.

Track 2 – CLDR as a Service

This presentation documents my initial attempt to create an open source, fully-containerized, CLDR web service that allows consumers of the web service to specify the CLDR edition, module, file and locales for which they want data, while filtering this data down to just the display names, currencies, units, numbering systems, and so on they need for a given application. The web service is being built in Typescript, runs on Node and leverages the CLDR npm packages developed by the Unicode Consortium. Presentation attendees will have the opportunity to work with the CLDR service themselves in the browser, and will learn about the decisions, processes and resources that went into its construction. By the end of the presentation, attendees will have a better understanding of what CLDR is, how CLDR is structured and how to access CLDR data through this new API.

Presenters:

Dr. Anshuman Pandey
Research Associate, SEI, Dept. of Linguistics, UC Berkeley

Dr. Deborah Anderson
Project Lead SEI/Researcher, Dept. of Linguistics, UC Berkeley

Track 3 – Historical Scripts and Innovations to Unicode Encoding Models

Historical scripts are often viewed as “secondary citizens” in the world of character encoding and internationalization. Modern scripts have large, living communities pressing urgently for their adoption in Unicode and implementation on devices. On the other hand, historical scripts do not have living informants; they rely on scholars and students to keep their usage and knowledge alive. Still, historical scripts are an important aspect of the Unicode Standard as they are bearers of the historical record of mankind. Encoding and implementing these scripts is the first step in building digital access to this textual record, which in many cases is threatened not only by the passing of time, but by natural calamities and human action.

Historical scripts are also important to the Unicode Standard for technical reasons. They have necessitated the development of specialized formatting controls for character positioning and text composition, such as for Egyptian Hieroglyphs. They have raised issues about ways to handle the representation of contextual elements, such as damaged text, partial forms, and deletions. Ancient scripts have also spurred discussions about the Unicode character-glyph model with regard to variant and alternate forms. The encoding of historical scripts has led to overall enhancements for complex layout and rendering and paved the way for encoding scripts across all time periods.

This talk addresses the innovations and challenges of historical scripts by describing the ongoing efforts to encode Egyptian Hieroglyphs, Maya Hieroglyphs, Proto-Sinaitic, Sidetic, and the Tocharian and Turkestani forms of Brahmi. The discussion will pose questions about how to handle the huge proliferation of signs during the evolution of a script, such as the numerous character variations in Egyptian Hieroglyphs from the Ptolemaic period, for which small distinctions between characters may express important semantic distinctions. It will raise issues about how to handle sign fragments and damaged text, such as in Egyptian Hieroglyphs, Proto-Sinatic, and Sidetic, which are important for preservation of textual corpora and epigraphical scholarship. The talk will also describe the need for expanding the conventional Indic model to handle syllabic conjuncts used in Central Asian forms of Brahmi. Lastly, the discussion will cover how restrictions on access to materials hinders the encoding process, as indicated by the difficulty in getting sources for Maya Hieroglyphs across wide geographical and temporal frames.

11:30-12:20 SESSION 2

Presenters:

Zibi Braniecki
Platform Internationalization TLM, Mozilla

Shane Carr
Senior Software Engineer, Google, Inc.

Nebojša Ćirić
Internationalization Team TL/M,
Google, Inc.

Track 1 – ICU4X: Advancing Modular Unicode Components

ICU4X is a new Unicode Consortium effort bringing i18n to modern platforms and devices.

Conceived at IUC 43, introduced at IUC 44, and now approaching public beta, ICU4X delivers ECMA-402 APIs, dynamic locale data, and ow resource usage in a composable, modular way using Rust, with interfaces to other programming languages. In this presentation, we will introduce the design principles of ICU4X, how to deliver locale data, and how to use it in multiple programming languages.

We will show how to take advantage of ICU4X's modularity to reduce binary size, and how ICU4X stacks up on memory and performance benchmarks.

Presenter:

Mark Davis
Lead Internationalization Architect, Google

Track 2 – Inflection Points

Many languages cannot be correctly written (or spoken) without grammatical inflections, where words change form based on grammatical features. These include grammatical number (plurality), gender, and case. CLDR and ICU have offered plural support for some years, allowing translated messages to accommodate a change in number. More recently, CLDR and ICU have added support for inflected units of measurement, such as kilograms or feet. This allows an implementation to format meters in Polish as “3 metry” (nominative) or “3 metrach” (dative) — the singular nominative form would be “1 metr”. In addition, the gender of the result is returned, allowing for implementations to adjust words in the surrounding message if necessary. There are two other important features:

- An enhancement of the structure for determining which units are used in which locales, allowing for choice of units by usage (measuring roads vs person-heights) and by size (meters below a certain distance; kilometers above).
- Computing grammatical gender, case, and number for compound units based on their components (such as furlongs per square fortnight).

This presentation will discuss the usage of grammatical features, show a demo of how the translations are gathered, and how the API works. It will also provide a short peek underneath the covers: how the grammatical feature data was gathered, how the underlying CLDR data is structured, and how the internal ICU code works. And finally, how to ramp up to more languages.

Presenters:

Norbert Lindenberg
Internationalization solutions Developer, Lindenberg Software LLC

Aditya Bayu Perdana,
Typographer of Indonesian Scripts

Track 3 – Learnings from Encoding Kawi

Kawi is the 64th Brahmic script encoded in Unicode, so one might think there's a semi-automated process in place that converts a description of a script into complete Unicode data including code points, combining character classes, Indic syllabic and positional categories, and line breaking classes. That's not the case. To some extent that's because each script has unique features that need special consideration, to some extent because the needs of different user groups such as scholars and online communities have to be balanced, to some extent because choices made for Brahmic scripts encoded earlier are no longer considered appropriate, and to some extent because the technical environment has changed.

This presentation looks at a number of decisions made for Kawi and their rationales and tries to turn them into recommendations for the next dozen Brahmic scripts. It also looks at the prototype implementation, which provides a Kawi font and keyboard for iOS, and gives experts an opportunity to validate the encoding before it's frozen.

12:30-13:30 - LUNCH

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

Presenters:

Zibi Braniecki
Platform Internationalization TLM, Mozilla

Eemeli Aro
Staff Software Engineer, Mozilla

Mihai Nita
Senior Software Engineer, Google, LLC

Track 1 – MessageFormat 2.0 – Localization for the Web and Beyond

Over the last several years, there has been a renewed energy around software localization coming from an increasingly multilingual user base. 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.

For the past two years, a Unicode working group has been collaborating on a solution to these problems by developing a new MessageFormat 2.0 standard. This standard aims to provide a single, uniform way for organizations to implement natural sounding messages, drawing on experience from past projects such as ICU MessageFormat 1.0, Fluent, FBT, Siri, and I18Next.

In this presentation, we will describe the Message Format 2.0 design principles, data model, syntax decisions, and target feature set. We will also present a prototype of MessageFormat 2.0 in action.

Presenter:

Andrew Poblocki
Software Architect, Salesforce

Track 2 – Extending International Formatting and Parsing Capabilities with @salesforce/i18n-Service-Library

The ECMAScript's Internationalization API Specification still has gaps in parsing and formatting. To deliver the full functionality that Salesforce customers require, we built @salesforce/i18n-service library. Learn how the library extends existing built-in functionalities with parsing, additional formats (ISO), and conversion between CLDR and Intl format specifications.

Presenters:

Craig Cornelius
Senior Software Engineer, Google, Inc.

Muhammad Noor
Co-Founder and Managing Director, Rohingya Project

Track 3 -Rohingya in Unicode – 3 Years After Standardization

The addition of Hanifi Rohingya script to Unicode has enabled numerous opportunities for the Rohingya people to use their language online and with mobile devices. Key supporting technologies including Unicode libraries, fonts installed on mobile devices, and keyboards that allow users to create and view content online with desk/laptop devices. Success and obstacles to full digital support are outlined.

This talk will review advances in Rohingya language support since the 2018 addition of this script to Unicode. The authors will also outline current initiatives and ongoing needs for full support for this community and its language, even while most of the speakers are in diaspora across many countries. For example, a Rohingya translated version of the universal declaration for human rights (UDHR) has been added to its list of World Language Translations (https://www.unicode.org/udhr/translations.html.)

The talk will also illustrate online and desktop applications and platforms that do not yet support the Rohingya script and fonts, including limitations on using the character set in documents as well as supporting personal names and other information in the script. These are significant obstacles for content developers and users. Finally, the authors will invite collaboration to fully include the Rohingya language in all platforms and products.

14:30-15:20 SESSION 4

Presenters:

Shane Carr
Senior Software Engineer, Google, Inc.

Greg Tatum
Staff Software Engineer, Mozilla

Track 1 – ECMA-402: i18n on the Web and Beyond

JavaScript remains one of the most widely used programming languages, and its popularity continues to grow. Today, it is used not only in web browsers, but also on mobile devices, IoT, and cloud services. This is why building great i18n into the ECMAScript standard library is a unique opportunity to have a high impact on the industry.

ECMA-402, the Intl object for ECMAScript, is standardized by TC39-TG2. In this presentation, delegates on TC39-TG2 will introduce the latest updates on ECMA-402 and how you can leverage it in your applications. We will cover the latest options for the formatting of dates, times, numbers, currencies, units, and durations; Unicode collation and segmentation; display names; plural rules; and what is next for the standards body.

We encourage attendees to visit github.com/tc39/ecma402 for information on how to get more involved with TC39-TG2.

Presenters:

Steven Loomis
Senior Software Engineer

Marcus Scherer
Unicode Software Engineer, Google, LLC

Track 2 – New in ICU & 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 14.0 & Emoji 14.0.

Presenters:

Dr. Anshuman Pandey
Research Associate, SEI, Dept. of Linguistics, UC Berkeley

Dr. Deborah Anderson
Project Lead SEI/Researcher, Dept. of Linguistics, UC Berkeley

Track 3 – Negotiating Neographies in Unicode: Approaches for Encoding Newly-Invented Scripts

While the majority of scripts in Unicode have long histories, the Standard also includes several modern scripts that were created in the 20th and 21st centuries. These new scripts, or ‘neographies’, were invented to provide the first indigenous script for a language; or to express unique linguistic and cultural identities, and to differentiate a minority community from surrounding predominant linguistic and orthographic cultures.
The majority of neographies have been invented in Africa and South Asia. In Africa, N'Ko and Adlam are examples of successful neographies, which have in turn motivated other linguistic and cultural groups to create their own scripts. In South Asia, Ol Chiki, Sora Sompeng, and Warang Citi are successful invented scripts that serve as ideological inspiration for new scripts. The aforementioned scripts have been encoded in Unicode, along with other neographies such as Masaram Gondi, Hanifi Rohingya, Wancho, and Toto. Tangsa and Nag Singh Mundari have also been successfully proposed for encoding. However, there are numerous scripts such as Tolong Siki, Jenticha Sunuwar, and Gurung Khema Phri, that have been proposed, but not yet encoded. As the number of neographies continues to grow in tandem with increased global awareness of Unicode, there has been an associated increase in proposals for encoding these new scripts.

The phenomenon of neographies offers interesting challenges for The Unicode Standard and for character encoding practices, in general. How to determine when a newly-invented script is ready for standardization in Unicode? What metrics can be used to understand the usage and stability of such scripts? What are the policy and cultural implications of a Unicode encoding for neographies?

In this talk, we discuss these questions and address criteria for evaluating proposals for new scripts by highlighting three categories: 1) brand-new scripts that are still gaining currency; 2) “near modern” scripts with relatively few users, but with a documented corpus; and 3) new scripts whose repertoires do not have consensus among the user community. For Africa, we present the cases of “Isibheqe Sohlamvu”, “Garay” and “Beria”, and for South Asia, “Tolong Siki”, “Jenticha Sunuwar”, and “Tani”.

15:20-15:50 - Afternoon Refreshments

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

Presenter:

TBA

Track 1 – TBA

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

Moderator:

Marcus Scherer
Unicode Software Engineer, Google LLC

Panelists:

Mark Davis
Lead Internationalization Architect, Google, Inc.

Shane Carr
Senior Software Engineer, Google, Inc.

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 will allow plenty of time for questions from the floor, and general Q&A.

Track 3 – After Standardization: How Do We "Really" Use Our Language Online?

Presenter:

Craig Cornelius
Senior Software Engineer, Google, Inc.

Unicode Standardization is a long process that gives hope to a language community for eventually becoming part of the online world. However, practical barriers remain after the Unicode Technical Committee completes its work. Expected benefits of a Unicode script are often inhibited by such limitations. This talk describes a set of open source tools and how they have helped groups come online with their languages.

For example, many languages use characters not readily available on keyboards and other input methods in provided layouts. Proposed layouts require evaluation for use on physical and soft screen devices. Fonts for the script may not be available for the script or may not reflect user expectations and preferences. And commercial fonts often render combining sequences incompletely or incorrectly.

Pre-Unicode text, defined with font-encodings or private use area (PUA), is an important resource that should not be lost. However, even conversions require building custom technical tools for file formats including word processing, spreadsheets, and presentations. In some cases, existing text in a dominant script such as Latin maybe transliterated into a newly standardized script such as Adlam and Lahkum (Tangsa). But the required tools are simply not available.

To address such needs, I built several AppEngine-based websites in collaboration with community language advocates and language experts. Using publicly available fonts and open source tools, the tool sites include adlamtesting.appspot.com, http://osagelanguagetools.appspot.com/, and \zawgyi-unicode-test.appspot.com/. The approach was extended to additional languages and writing systems on https://languagetools-153419.appspot.com/, using an easily extended common approach for scripts and languages while supporting customization for specific requirements and exploration. These are implemented via server-side Python with Django templating and Javascript client code. The code is publicly available on GitHub https://github.com/sven-oly.

The tools include these general options

- Prototypes of keyboard layouts and transform rules applying Unicode fonts in addition to “Unicodifed” custom fonts conversion from existing font-encoded text to Unicode
- combining mark combinations downloadable resources and links to additional language data
- Additional tools exist for collation and building multilingual lists as needed
- New capabilities are under development, including implementing CLDR keyboard formats

The presentation will outline the needs and motivations for creating these tools and will give examples of their use. The author welcomes comments, contributions, and collaborations in the hope of aiding communities worldwide to conveniently and effectively use their languages digitally on all platforms and applications.

16:50-17:40 SESSION 6

Presenter:

Addison Phillips
Senior Principal Engineer, Amazon.com

Track 1 – Time Out of Joint: Working with Dates, Times and Time Zones in Java

How can you accurately tell your customer when their package will be delivered, their favorite TV show will be available, or when their subscription will renew? Customers expect precise and meaningful date and time values. Providing these to customers can be difficult if you don't understand the use cases you're managing and the complexities of time zone. This presentation lays out the basics using the Java programming language and shows how solutions such as java.time (Joda) and ICU4J can be used to address common scenarios.

Moderator:

Stephen Loomis
Senior Software Engineer

Panelists:

Joel Sahleen
Globalization Architect, Domo, Inc.

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 the 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. Topics will include implementation hurdles when using CLDR from JSON.

Presenters:

Craig Cornelius
Senior Software Engineer, Google, Inc.

Robert Melo
Software Engineer, Apple

Track 3 – Enabling Latin American Indigenous Languages for Android & Unicode CLDR

Studies say that we had approximately 1200 languages before Europeans arrived in Brazil. After 500 years, about one thousand languages do not exist anymore. At the same time, researchers state the surviving languages may not exist in the next decades. What can we do to avoid the disappearance not only of the languages, but of histories and cultures?

This presentation reflects an Internationalization initiative at Motorola centered in the effort of development of written usage of indigenous languages. When new generations of indigenous people become part of a literate world, the integration of their native languages in the written world is crucial to the survival of their languages and cultures. In parallel, indigenous languages need to be present everywhere to be seen and heard, even for those who do not understand them, and the Internet is an obligatory space for all languages.

Based on that, the talk will highlight the efforts and requirements necessary to enable a language on Android & CLDR, standardize orthography, work with technology platforms and vendors, adapt existing resources (e.g. operational systems, keyboards, apps, etc.). Join us to discuss how we can contribute to the addition of more indigenous languages and be prepared for the decade of indigenous languages that starts next year. Learn about language revitalization as well as the joys and challenges of developing the Unicode proposal.

18:00-19:00 – CONFERENCE NETWORKING RECEPTION

Friday, October 15, 2021

09:00-09:50	SESSION 7

Presenters:

Filip Filmar
Software Engineer, Google, LLC

Konstantin Pozin
Software Engineer, Google, LLC

Track 1 – rust_icu: a Rust Language Binding for ICU4C

As the Rust programming language matures into a tool of choice for building robust system level applications, there is an increased need to handle Unicode text processing as well. While several efforts to build Unicode support for scratch are ongoing, ICU4C is already there. rust_icu brings ICU4C to Rust today. We developed rust_icu to answer the specific requirements of the Fuchsia OS but recognized early on that it may be useful beyond that narrow domain, so we decided to offer it as an open source library.

Presenter:

Mike McKenna
Director World Ready Engineering, PayPal, Inc.

Track 2 – CLDR and Person Names – The Saga Continues . . .

Last year, we presented at ICU44 the problems and many solutions that have been created for how to handle personal names. We have spent the intervening time discussing comments that have arisen and refining how personal name metadata should be structured to fit well in the CLDR universe and be straight-forward to implement in ICU, intl.js, or other platforms, with a goal for it to be intuitive to the developer when creating locale-sensitive interfaces.

The proposed Personal Name Structure has been submitted to the Unicode CLDR committee as a PRI (Public Review Issue) for the CLDR v40 cycle. It includes a data model, fall back mechanisms, example "skeletons" used to construct various name layouts depending on the chosen content, and API examples. The proposal also discusses what is in scope and what is out of scope. E.g. CLDR will not encode linguistic or grammatical rules in its first rendition. This session will present the specifics of the proposal as submitted and walk-through examples and use cases for how it is envisioned to be used in real life.

Presenters:

Craig Cornelius
Senior Software Engineer, Google, Inc.

Mark Durdin
Keyman Team Lead, SIL International

Andrew Glass
Senior Program Manager, Microsoft

Joshua Horton
Keyman Predictive Text Lead, SIL International

Stephen Loomis
Senior Software Engineer

Track 3 – Standardizing Keyboards with CLDR

More and more language communities are determining that digitization is vital to their approach to language preservation and that engagement with Unicode is essential to becoming fully digitized. For many of these communities, however, getting new characters or a new script added to The Unicode Standard is not the end of their journey. The next, often more challenging stage is to get device makers, operating systems, apps and services to implement the script requirements that Unicode has just added to support their language.

In recent years, the standardization of complex font shaping ([USE](https://aka.ms/universalshapingengine), [Harfbuzz](https://github.com/harfbuzz)) and Google's Noto Fonts have contributed significantly to reducing the barriers and shortening the timeline for the correct display of additions to Unicode. However, commensurate improvements to streamline new language support on the input side have been lacking. CLDR's new Keyboard Subcommittee has been established to address this very gap.

This presentation will introduce the work of the Keyboard Subcommittee, and the roadmap ahead to standardize keyboard layout data. The end results of this effort are projected to be:

- A comprehensive standard ([LDML](https://www.unicode.org/reports/tr35/tr35-keyboards.html)) that defines keyboard layout data (hardware layouts, virtual layouts, and deterministic transforms)
- An updated version of [Keyman](https://keyman.com/) that will enable language communities to develop keyboard layouts in the LDML standard
- A single hub ([CLDR](http://cldr.unicode.org/)) for definitive keyboard layout details in LDML format which can be easily consumed by any endpoint that offers input support.

The results of this broad effort promise to significantly reduce the costs and complexity of supporting new languages for input and thereby speed up the time to implementation for newly digitized languages.

10:00-10:50	SESSION 8

Presenters:

Mathias Bynens
JavaScript Whisperer, Google, LLC

Mark Davis
Lead Internationalization Architect, Google, Inc.

Marcus Scherer
Unicode Software Engineer, Google LLC

Track 1 – Unicode Regular Expressions for JavaScript

We will present and discuss new developments in the Unicode regular expressions standard (UTS #18) and related proposals for improvements in JavaScript (ECMAScript) regular expressions.

Driven by developer demand for detecting and working with emoji in web apps, the UTC has added definitions for "properties of strings" and clarified how those interact with regular expression "character classes". We are working with Ecma TC39 on proposals to bring several emoji properties to JavaScript developers, as well as set operator syntax to make it possible and easy to build character classes based on properties but with additions and exceptions, which are often necessary. We are aiming for developers to be able to write concise, readable, self-updating regular expressions, replacing many kilobytes of generated character class data or custom code.

More generally, increasing support for multi-character strings will also help remove hurdles for handling “characters” that are encoded as sequences of code points.

Presenter:

Jim DeLaHunt
Principal, Jim DeLaHunt and Associates

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

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://普遍适用测试.我爱你 to تجربة-بريد-الكتروني@تجربة-القبول-الشامل.موريتانيا to सार्वभौमिक-स्वीकृति-परीक्षण.संगठन . We explain the most troubling obstacles and the most inspiring successes in Universal Acceptance encountered by the Universal Acceptance Steering Group. From major email platforms launching support of internationalized addresses to improving support by programming languages and libraries, it has been an exciting year.

Presenters:

Mark Durdin
Keyman Team Lead,
SIL International

Joshua Horton
Keyman Predictive Text Lead,
SIL International

Track 3 – Community-Sourced Predictive Text in Keyman

We will examine the relationship between keyboards and text correction on touch-based devices and walk through the steps required to start providing predictive text support for any indigenous language through use of the Keyman engine. We'll also examine some of the optional features for models that can be used to better tailor the engine’s behavior to better fit the expectations of a model's target language community.

As technology is playing an increasingly large role in the everyday lives of people all over the world, facilitating the use of indigenous and endangered languages, especially on consumer-friendly end-products such as mobile devices, will aid language preservation and revitalization. Removing technological roadblocks for use of Indigenous and endangered languages helps preserve their use in day-to-day life as communication shifts increasingly to electronic-based media, even in small and remote communities.

Keyman, an open-source input method platform originally designed in 1993, has long had support for such language communities in mind. Over the past few years, through a partnership with the National Research Council of Canada, we have also begun incorporating a predictive text engine written in TypeScript into our apps for Android and iOS mobile devices. Our design goal is to dramatically lower the requirements needed by a language community for predictive text, allowing them to quickly kickstart predictive text support for their language, making it available to their language communities as rapidly as possible.

10:50-11:10 - Morning Refreshments

11:10-12:00 SESSION 9

Moderator:

Alolita Sharma
Director, Unicode Consortium

Track 1 – The Pandora's Box BoD/Town Hall/Issues

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.)

Presenters:

Shibi Sudhakaran
Senior Architect, PayPal

Neha Utkur
Engineering Manager, PayPal

Track 2 – Building Global Products @ PayPal

PayPal is a truly global company. Its products are available in more than 200 markets and support over 40 languages. Large back-end systems plus modular applications that are being continuously released need to conform to internationalization standards from CLDR, ISO, IANA, and other industry sources, as well as various regional standards.

Product development teams have had to chase down the data and interfaces to all this standardized regional information from different teams, different repositories, and sometimes different continents. Because of the global size of PayPal, multiple teams have reinvented the wheel to use these data in their products.

Enter World Ready SDK (i18n library) and World Ready Service! By providing a holistic framework encompassing i18n client libraries and a highly scalable, available and extensible service, the framework provides regional information, compiled and massaged into a SaaS solution, based off standardized regional information from CLDR, ISO, IANA, and other industry sources to enable applications to dynamically configure themselves without having to load a single large set of data. On top of that, by decoupling content resource bundles from the application codebase, this framework also reduced the content turnaround time to live sites from weeks to hours!

In this session, we'll go over the architecture of the unified framework as well as the challenges and pitfalls we ran into while building the solution. We will conclude with how product and development teams at PayPal have benefitted from this effort.

Presenter:

Mike McKenna
Director World Ready Engineering
PayPal, Inc.

Track 3 – Locale and Context Sensitive Input Mechanisms

Creating a dynamic interface for user input of personal information is not always easy. You ask them to enter their name, address, phone number. But first you need to know what region they are in. Then what language they wish to use for the user experience.

Those are the easy parts. What if they are entering an address and phone for someone else, in another region? What language should the labels be in? If they are working in a Thai interface, and start entering an address in Latin/English, how do you detect that? And if so, do you then start showing the input pull-downs in English, even though the rest of the UI is in Thai? Then ... what language and formatting do you use for messages and info from a user in one language and region to another person in another region and language?

This session will walk through the design decisions made to answer these questions, then walk through a sample application to see how it dynamically changes its layout and content depending on the session context characteristics and user-entered-data using Unicode character properties, CLDR best-guess locales, and extensible object design.

12:00-13:00 - LUNCH

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

Presenters:

Peter Constable
Sr. Program Manager for Globalization and Typography Standards, Microsoft

Dominik Röttsches
Staff Software Engineer – Chrome, Google, LLC

Rod Sheeter
Software Engineer – Google Fonts, Google, LLC

Track 1 – Vector Color Fonts

Color fonts have historically been hard to deliver to the web due to 1) many color font formats, none widely supported in browsers and 2) lack of a format with rich capabilities that is also web-friendly (compression-friendly). As a result, web users tend to rely on techniques like scanning for emoji sequences and inserting images. We will discuss advancement in the following areas:

- Font specification updates
- (https://github.com/googlefonts/colr-gradients-spec)
- Input layouts, display layouts, - and validations
- Open source tools for building color fonts
- (https://github.com/googlefonts/nanoemoji)
- Compressed and decompressed file sizes for a range of color formats
- Color fonts on the Web

Presenter:

Samuel Parmar
Senior Engineering Manager, Salesforce

Track 2 – Evolution of Translation Use-Cases

With rapid digitization across different industries, your customers' expectations around localization are changing quickly. Traditionally, only static texts, labels, and error messages are translated to end-user languages. However, there are many emerging use cases where customers want to translate more dynamic data to enhance localization experience, such as Chatbot.

This talk covers Salesforce's Data Translation feature and how it extends the Salesforce Platform capabilities, allowing customers to decide which data they want to translate and how to build the customer experience around localization.

Presenters:

Mark Davis
Lead Internationalization Architect, Google, Inc.

Luke Swartz
Product Manager, Google, LLC

Daan Van Esch
Technical program Manager, Google, LLC

Track 3 – Does Device X Support My Language?

People often ask whether some device or application supports their language. This seems like a simple question: yes or no. But the reality is that there are different levels of support for a language, ranging from only allowing the user to read their language on the platform all the way up to having the platform fully localized in their language, with features such as a voice assistant.

This presentation proposes a common set of terminology for language support levels for platforms such as operating systems, browsers, etc. The goal is to have consistent terminology so that people can clearly understand what the possible levels of support for a given language are. In addition, it presents guidance for how to improve the level of support for a given language: what concrete steps can be taken.

13:50-14:40 SESSION 11

Presenters:

Arlene Ducao
Founder, Adjunct Professor, Multimer, NYU &I CUNY

Bex Hurwitz
Founder, Tiny Gigantic

Track 1 – Emojis and Inclusion

Emojis have become a part of our digital language. We use emojis to relay messages, express emotions, sling slang. Not everyone knows that anyone can submit a proposal to the Unicode Consortium's Emoji Subcommittee for a new emoji, and that only a few are accepted.

In this presentation, we'll discuss the findings from a series of "Emoji as Digital Language" workshops we have prepared for diverse students at universities in the United States and Canada. These workshops are meant to highlight the roles emojis play in participants' lives, communication, and meaning-making. They also help participants learn about the Unicode Consortium, internet standards, and ask questions about who proposes emojis, judges' proposals, and determines what becomes an emoji. Participants are meant to leave with practical information about how to propose emojis and deeper understanding of internet standards and how a more inclusive internet can be encoded.

Presenter:

Tex Texin
Chief Globalization Architect, Xencraft

Track 2 – Web Internationalization

This former tutorial, updated in 2021, 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 session 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.

Presenters:

Craig Cornelius
Sr. Software Engineer, Google, Inc.

Jeannette Stewart
Founder, Translation Commons

Track 3 – Accelerating Support for Indigenous Languages in Digital Systems

This presentation will focus on creating awareness of the need for language digitization and why accelerating implementations has now become vital to Indigenous communities. UNESCO will be highlighting this need during the International Decade of Indigenous Languages which begins in 2022. The Los Pinos Declaration, signed at a 2020 UNESCO event, places indigenous peoples at the centre of its recommendations under the slogan “Nothing for us without us.” The Declaration, designed to inspire a global plan of action for the Decade, amongst many directives, also points to the potential of digital technologies in supporting the use and preservation of those languages. We need to understand the different circumstances of indigenous peoples in order to empower them to seek language technology solutions for their communities.

Many people read and write only in languages that have no digital support. With limited access to information critical to modern living, they cannot participate in e-commerce, e-mail, searching the Web for information on healthcare, how-to articles, etc. They miss critical information from their government and timely instructions when a disaster strikes. They cannot contribute to the Web, social or print media in the many ways routinely available to the speakers of supported languages to add knowledge or point of view to society in general, to government representatives, etc.

The session will also highlight several case studies of this process for indigenous communities, including the obstacles and challenges as well as positive steps toward digital language capability. Examples range from individuals such as Bivuti Chakma digitizing single-handedly the Chakma language, all the way to entire communities uniting together to bring their language online such as the Sunuwar community in Nepal and Sikkim. Specifically, we will examine the Sunuwar language community attempting their first steps in creating a Unicode Proposal for their script encoding, and we will identify the struggles and challenges faced by indigenous peoples wanting to use their language digitally.

Translation Commons first created the Language Digitization Initiative for communities who want the above benefits of language digitization. It defines a process to move a language through the steps of digitization. Important initial steps are Unicode standardization, creating keyboard systems, defining fonts, and making these available and usable for community members. Now in partnership with UNESCO's International Year and Decade of Indigenous Languages, Translation Commons is preparing the Language Technology Framework to be implemented in the upcoming Decade.

14:50 – 15:10 - Afternoon Refreshments

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

Presenter:

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

Track 1 – Talk Emoji to Me

When there are as many foods as there are ingredients on the planet, and a variety of objects only limited by your imagination, every addition to the emoji palette is at risk of creating zones of exclusion without consciously trying. A look at how the ESC reconciles the rapid transient nature of communication with the formal methodical process of a standards body like Unicode 🎷🐛

Presenter:

Claudia Galván
Technical Advisor, Early Stage Innovation

Track 2 – Measuring Success with Internationalization KPI's

Key Performance Indicators help drive decisions to increase, performance, revenue and international strategy. This talk will cover understanding the customer metrics, operational metrics and business metrics used to measure success in the international markets.

Presenters:

Julie Anderson
Advisor, Translation Commons

Craig Cornelius
Sr. Software Engineer, Google, Inc.

Track 3 – Digitization Solutions for Indigenous Languages

Out of 7,000 languages spoken in the world today, only several hundred or so are fully supported for digital use, leaving the remaining language communities at a significant disadvantage. In this session you will learn about the Language Digitization Initiative (LDI) and how it is achieving its goal for a world in which all language communities have equal digital opportunities. There are many components to this initiative, including the enabling of indigenous communities to pursue and manage the digitization process themselves.

Julie Anderson will give a brief introduction to the initiative and Craig Cornelius will describe the following guidelines being published this year, with Indigenous communities as their intended audience: Language Data Gathering, Language Data Repositories, Terminology Management, and Machine Translation.

These are part of our series "Zero to Digital" which addresses all of the steps in the language digitization process. The guidelines are authored by experts in language technology and linguistics, many of whom are known to the Unicode community.

The process of gathering language data is necessary for defining basic keyboard, font, and script characteristics. Repositories are needed to host and organize the data and allow for sharing among researchers, linguists, and for community contribution, review, license rights, and privacy controls. Terminology management systems enable documentation of definitions of terms, grammar usage, and translations. Terminology information is used by digital systems for spell and grammar checking, predictive text, Optical Character Recognition (OCR), Automated Speech Recognition (ASR), and many other functions. At the very advanced end of digitization, and if a large enough language corpus exists, machine translation may become feasible.

Digitization of minority languages is significant for these communities, as it democratizes access to information, helps address socioeconomic inequalities, promotes emerging markets, sustains minority languages, and preserves priceless cultural heritage.

The LDI program was created by Translation Commons (TC), a nonprofit volunteer community of language professionals. Its programs sustain endangered languages and cultures, support volunteers in their professional development, and disseminate linguistic knowledge and awareness.

16:10 - 17:00 CLOSING SESSION

Moderators:

Martin Dürst
Professor, Aoyama Gakuin University

Alolita Sharma
Principal Technologist, AWS

Lightning Talks

This traditional closing session will be a series of lightning talks of 5-10 minutes each, followed by extremely short closing remarks. The talks should be related to internationalization, localization, or any other of the topic areas listed in the Call for Participation. This is the chance for you as a conference attendee to present your latest idea or development, spread the word, or raise awareness about something of importance to you, or talk about a topic that doesn't need a full session, or a conclusion or question you are taking home from the conference. Talks should be 5 minutes in length maximum. If interested, please send an email to the Moderators, Martin Dürst (duerst@it.aoyama.ac.jp) and/or Alolita Sharma (alolita.sharma@gmail.com). Questions on any of the lightning talks will be at the end of the session.