Program - Session Descriptions

Monday, October 17, 2011

09:00-12:30 MORNING TUTORIALS

Presenter:

Richard Ishida
Internationalization Lead, W3C

Track 1: An Introduction to Writing Systems & Unicode

The tutorial will provide you with a good understanding of the many unique characteristics of non-Latin writing systems, and illustrate the problems involved in implementing such scripts in products. It does not provide detailed coding advice, but does provide the essential background information you need to understand the fundamental issues related to Unicode deployment, across a wide range of scripts. It has also proved to be an excellent orientation for newcomers to the conference, providing the background needed to assist understanding of the other talks! The tutorial goes beyond encoding issues to discuss characteristics related to 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. The concepts are introduced through the use of examples from Chinese, Japanese, Korean, Arabic, Hebrew, Thai, Hindi/Tamil, Russian and Greek. While the tutorial is perfectly accessible to beginners, it has also attracted very good reviews from people at an intermediate and advanced level, due to the breadth of scripts discussed. No prior knowledge is needed.

Presenter:

Addison Phillips
Globalization Architect Lab126 (Amazon)

Track 2: Internationalization: An Introduction, Part I: Characters and Character Encodings

What is internationalization? What do developers, product managers, or quality engineers need to know about it? How does a software development organization incorporate internationalization into the design, implementation, and delivery of an application?

This tutorial track provides an introduction to the topics of internationalization, localization and globalization. Attendees will understand the overall concepts and approach necessary to analyze a product for internationalization issues, develop a design or approach, and deliver a global-ready solution. The focus is on architectural approaches and general concepts, but will include specific examples and exercises.

Part I focuses on characters, character encodings, and the basics of Unicode.

Presenter:

Thomas Milo
Partner, DecoType

Track 3: Comprehensive Arabic Script Tutorial

This is a completely revised and updated comprehensive tutorial presented many times before to the Unicode Conference covering all aspects of Arabic-script computing, from calligraphy to typography history, script structure, orthography, encoding, dumb and smart computer typography, types of line-breaking, language specific issues, and much more.

10:30-10:45 - Morning Refreshments

Presenter:

Richard Ishida
Internationalization Lead, W3C

Track 1: An Introduction to Writing Systems & Unicode (Cont'd.)

The tutorial will provide you with a good understanding of the many unique characteristics of non-Latin writing systems, and illustrate the problems involved in implementing such scripts in products. It does not provide detailed coding advice, but does provide the essential background information you need to understand the fundamental issues related to Unicode deployment, across a wide range of scripts. It has also proved to be an excellent orientation for newcomers to the conference, providing the background needed to assist understanding of the other talks! The tutorial goes beyond encoding issues to discuss characteristics related to 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. The concepts are introduced through the use of examples from Chinese, Japanese, Korean, Arabic, Hebrew, Thai, Hindi/Tamil, Russian and Greek. While the tutorial is perfectly accessible to beginners, it has also attracted very good reviews from people at an intermediate and advanced level, due to the breadth of scripts discussed. No prior knowledge is needed.

Presenter:

Addison Phillips
Globalization Architect Lab126 (Amazon)

Track 2: Internationalization: An Introduction (Part II, Writing Global Ready Code)

Part II focuses on preparing for the localization (translation) of user interfaces; making applications "locale-aware", including format and display differences; as well as approaches to delivering multi-lingual and multi-locale software or content.

Presenter:

Su Liu
AIX Globalization Architect, IBM

Track 3: Smart Code Set Conversions for Unicode Support in Heterogeneous Environments

Modern storage network includes hundreds of code sets, and thousands of conversion modules for information services. In the heterogeneous environment, it is a crucial criterion to dynamically, efficiently, accurately convert Unicode data to non-Unicode data or vice versa. Therefore, the developing, and maintaining Unicode conversion services become more challenging tasks due to the inconsistent code set names, multiple encoding schemes, variants of code mapping tables, multiple versions of encoding standard, and diverse OS platforms. This tutorial introduces mechanism in the code set conversion design and explains use of Unicode technologies for solving problems in layers of application, operating system, and network. The tutorial first compares the differences among the major code set conversion algorithms, and then focuses on the challenges of Unicode and non-Unicode conversions in internal UNIX and across networks. A further discussion into the smart and advanced conversion introduces methods and solutions related to conflicts and problems on modifier, endianness, composed/decomposed character, multiple code set standard versions, and code set alias names. It gives some examples to illustrate options of complex text and CJK manipulations. Finally, the tutorial addresses the code set converting implementation strategies and options to choose native UNIX based or ICU based conversion functions for future Unicode support in the heterogeneous environment.

Keywords: Unicode, AIX, UNIX, Code Set, Conversion, Operating System

12:30-13:30 - LUNCH

13:30-15:30 **AFTERNOON TUTORIALS**

Presenters:
Craig R. Cummings

Michael G McKenna
Senior International Engineering Manager, Zynga

Track 1 - Unicode - A Grand Tour

This tutorial will cover the next level of detail of what Unicode is, and how it is used in the real world. The modules of the tutorial include: The Unicode standard - what are the "Guiding Lights", or design principles behind Unicode? A tour of Unicode's structure, encoding forms, behavior, technical reports, database, and how to use the Unicode Standard. Implementation according to Unicode - a walk through the details of attributes, compatibility, non-spacing characters, directionality, normalization, graphemes, complex scripts, surrogates, collation, regular expressions and other aspects according to the Unicode Standard and associated Technical Reports. Unicode and the Real World - an overview of Unicode-based internationalization development libraries and implementations supporting Unicode in web servers, application servers, browsers, C/C++, Java, PHP, SQL, and various operating systems. On-going programs - how Unicode is evolving and how you can participate in its future. Pointers to other sessions at the conference that dive deeper on particular topics are highlighted throughout.

Presenter:

Tex Texin
Chief Globalization Architect, Rearden Commerce, Inc.

Track 2 - Tutorial Web Internationalization - Standards and Best Practices

This tutorial is an introduction to internationalization on the World Wide Web. The audience will learn about the standards that provide for global interoperability and come away with an understanding of how to work with multilingual data on the Web. Character representation and the Unicode-based Reference Processing Model are described in detail. HTML, including HTML5, XHTML, XML (eXtensible Markup Language; for general markup), and CSS (Cascading Style Sheets; for styling information) are given particular emphasis. The tutorial addresses language identification and selection, character encoding models and negotiation, text presentation features, and more. The design and implementation of multilingual Web sites and localization considerations are also introduced.

Presenter:

Jim DeLaHunt
Principal Jim DeLaHunt & Associates

Track 3 - Building multilingual websites in Drupal 7 and Joomla 1.6

A practical look at the language and locale capabilities of Joomla! 1.6 and Drupal 7, two leading free software content management systems (CMSs). They let you build more powerful, more international websites faster. We look at: their core internationalisation and locale services; localisation of UI and content. Each platform just had a major release, with advances in internationalisation. You will leave with specific tips for building your own site. We don't assume Joomla or Drupal experience, but do include material for advanced practioners. A good tutorial for web site product managers, web designers, developers, and managers of international web teams.

15:30-15:45 - Afternoon Refreshments

15:45-17:45 **AFTERNOON TUTORIALS**

Presenters:

Michael G McKenna
Craig R Cummings *Senior International Engineering Manager, Zynga*

Track 1 - Unicode - A Grand Tour (Cont'd.)

This tutorial will cover the next level of detail of what Unicode is, and how it is used in the real world. The modules of the tutorial include: The Unicode standard - what are the "Guiding Lights", or design principles behind Unicode? A tour of Unicode's structure, encoding forms, behavior, technical reports, database, and how to use the Unicode Standard. Implementation according to Unicode - a walk through the details of attributes, compatibility, non-spacing characters, directionality, normalization, graphemes, complex scripts, surrogates, collation, regular expressions and other aspects according to the Unicode Standard and associated Technical Reports.

Unicode and the Real World - an overview of Unicode-based internationalization development libraries and implementations supporting Unicode in web servers, application servers, browsers, C/C++, Java, PHP, SQL, and various operating systems. On-going programs - how Unicode is evolving and how you can participate in its future. Pointers to other sessions at the conference that dive deeper on particular topics are highlighted throughout.

Presenter:

Track 2 - Internationalization Testing Best Practices

Loïc Dufresne de Virel
Localization Strategist
Intel

In this tutorial, attendees will learn how to develop a systematic I18N validation plan. Touching on code scans, pseudo-builds, pseudo-locales, testing on localized Operating Systems, and international test data, the authors will share their years of experience in the field of internationalization & localization, showing you how to identify I18N issues as early as possible to avoid taking on unexpected amounts of technical debt! To illustrate their presentation, they will use actual issues that were found, and sometimes missed, during recent localization projects - guaranteeing a fun and practical session.

Co-authors / presenters:
Michael Kuperstein, Intel, Localization Engineer
Octavio Ramos, Intel, Validation Lead

Presenter:

Track 3 - Using ICU Workshop

John Emmons
Senior Software Engineer
IBM

This tutorial gives attendees everything they need to know to get started with working with text in computer systems: character encoding systems, character sets, Unicode, and text processing, using the International Components for Unicode library (ICU).

ICU is a very popular internationalization software solution. However, while it vastly simplifies the internationalization of products, there is a learning curve.

The goal of this tutorial is to help new users of ICU install and use the library. Topics include: Installation (C++ libraries, Java .jar files, Java SPI for JDK integration), verification of installation, introduction and detailed usage analysis of ICU's frameworks (normalization, formatting, calendars, collation, transliteration). The tutorial will walk through code snippets and examples to illustrate the common usage models, followed by demonstration applications and discussion of core features and conventions, advanced techniques and how to obtain further information. It is helpful if participants are familiar with Java, C and C++ programming. Issues relating to ICU4C/C++ as well as ICU4J (Java) will be discussed. After the tutorial, participants should be able to install and use ICU for solving their internationalization problems.

18:00-19:00 - Welcome Reception

Tuesday, October 18, 2011

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

09:15-10:00 ***KEYNOTE PRESENTATION - Building the Multilingual Web - The Long View***

Presenter:

Laura Welcher

The Rosetta Project at The Long Now Foundation is working to build an open public digital collection of all human language as well as an analog backup that can last for thousands of years - The Rosetta Disk.

Director of Operations, The Rosetta Project

In the "long now," the goal is long-term storage and access to information on the scale that both supports and transcends individual human societies and civilizations. In the "here and now" the project serves to support and amplify the importance of the world's nearly 7,000 human languages, the vast majority of which are endangered and, if current trends continue, likely to go extinct in the next 100 years.

The Rosetta Project shares the Unicode vision of a world where people can use communication technology on their own terms - in their own language. According to World Internet Statistics, over 80% of all web communication is in about ten languages, with over half in either English or Chinese. The remaining 20% represent "everyone else" including about 400 languages with speaker populations above 1 million, which collectively comprise about 95% of everyone on earth. Because of essential technologies like Unicode, we are poised to see this breadth of human languages flourish online and on mobile devices, providing for these languages a critical new domain of language use in the modern world. I will present several efforts underway at The Rosetta Project including the "Language Commons" that rely on Unicode as an essential technology in building the multilingual Web.

10:00-20:00 - EXHIBIT AREA OPEN

10:00-10:30 - Morning Refreshments in Exhibit Area

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

Presenters:

Addison Phillips
Globalization Architect Lab126 (Amazon)
Richard Ishida
Chair, Activity Leave W3C Internationalization WG

Track 1 - To the Promised Land: I18N Developments in HTML5 and CSS3

A new era of competition between the major browsers has rekindled work on the HTML and CSS standards. One of the results has been a renewed focus on providing features to support international content in new and exciting ways, from typeset quality presentation to vertical text; from East Asian support such as ruby and emphasis to improved bidirectional language support.

In this presentation we'll explore the changes that are available today, the status of these standards, and the challenges that remain.

Presenter:

Aharon Lanin
Software Engineer, Google Inc.
Andy Staudacher
Software Engineer, Google Inc.
John Tamplin
Software Engineer, Google Inc.
Katsuhiko Momoi
Staff Test Engineer, Google Inc.
Mark Davis, Sr.
Internationalization Architect, Google Inc.

Track 2 - Pseudolocalization at Google -- some innovations

In this presentation we present some innovations on pseudo locale uses within Google with particular focus on standardizing pseudo locale naming scheme based on BCP 47, defining a small set of standard pseudo locales and their exact definitions, and how they can be used for automated checks for detecting major internationalization and bidirectionality issues. To spread good practices in internationalization testing across the industry, we recently released an open source pseudolocalization Java library. Standardizing pseudo locales is essential for this purpose as well.

We begin with a brief introduction to pseudolocalization concepts, its current merits and limitations. We argue that standardizing the naming scheme (BCP 47 compliant) and defining a few standard pseudo locales offers many benefits for Google's development environments and similar ones elsewhere.

A project can host other projects and expect to see across the components exactly the same pseudo locale methods used with the standard pseudo locales: "en-psaccent" (the LTR pseudo locale) and "ar-psbidi" (the FakeBidi pseudo locale). We will demo a BidiChecker tool that can run on the FakeBidi locale.

These innovations in the use of pseudo locales allow us to detect localizability issues early via automated checks, encourage development of new tools and tests on them, and help improve internationalization quality of products.

Presenter:

Track 3 - Implementing Better Source Editing for Bidirectional HTML and XML in the Text Editor Emacs

Martin J. Dürst
Profesor
Aoyama Gakuin University
Shunsuke Oshima
Master Course Student
Aoyama Gakuin University

Authors: Shunsuke Oshima and Martin J. Dürst

The Unicode Bidirectional Algorithm (UBA) is tailored for running text such as letters and newspaper articles. However, it is not suited directly for structured formats such as XML, HTML and programming languages. The source is often reordered in unpredictable ways that are unrelated to the logical structure of these formats, and therefore, source editing was essentially impossible. In this paper, we present a solution to this problem and its implementation in the text editor Emacs.

Emacs is a very flexible and extensible text editor providing an integrated environment for a wide variety of development tasks. Extensibility is based on Emacs Lisp, which we also have used for our research. While there have been some experimental implementations of bidirectional rendering in Emacs in the past, a full implementation of the UBA has only become available recently in Emacs version 24.

In earlier research, we implemented a Web-based simulation for XML and XHTML source rendering (IUC28) and a JavaScript-based experimental editor (IUC32). These implementations, however, were standalone and did not reach the level of practical usability.

The problem of bidirectional source editing for structured formats such as (X)HTML and XML can be divided into three areas. The first area is the treatment of syntactically significant characters, for example the ubiquitous angle brackets in HTML and XML. The UBA classifies them as neutral so that they follow the direction of their surrounding text, including potential mirroring, which is appropriate for running text. However, in HTML and XML they define the overall structure of the markup and therefore have to be treated as strong.

The second area is the treatment of bidirectional control characters such as LRM and RLM in source editing. Without intervention, such characters can be entered either literally, in which case they show their effect but are invisible and therefore difficult to edit, or they can be entered in escaped form (e.g. ‎ or ‏) in which case they are visible but not effective. Ideally, they would be both visible and effective.

The third area is bidirectional markup such as the 'dir' attribute or the <bdo> element in HTML, which for ease of editing should be reflected in the layout of the element content during source editing.

All the above areas can be addressed by carefully placing additional bidirectional control characters into the source text. We already had worked out much of the details of this placement in our earlier research. The main difficulty with using additional bidirectional control characters is that they are not part of the actual source text and therefore have to be distinguished from the same characters when they are part of the source, and have to be carefully removed for operations such as copying and saving. We implemented two different ways of doing this. One way uses a special Emacs property to distinguish these characters so that they can be removed before the relevant operations. The other way uses Emacs overlays, which by definition are not part of the text proper. In many ways, this would be the ideal solution, but overlays currently are not taken into account for bidirectional rendering.

Another problem is that the inserted bidirectional control characters have to be recalculated for every single editing operation. Limiting insertion to the currently visible part of the text being edited makes sure that we achieve acceptable

performance even for very long source files.

We are currently extending our implementation to work with TeX, and are looking into ways to fine-tune our implementation based on user feedback.

11:30-12:20	SESSION 2

Presenter:

Track 1 - Enterprise PHP Internationalization and Localization: A Case Study

Joel Sahleen
Software Developer, Adobe Systems Incorporated

PHP is one of the most popular server-side scripting languages on the internet. According to the latest data from builtwith.com, there are currently more than 25 million sites that run on PHP, including a third of the top million most visited. Although PHP may be best known for its use in high-profile, consumer-oriented sites such as Facebook and Yahoo!, in recent years the language has begun to carve out a new role for itself as part of the technology stack used to build business-critical, enterprise web applications like the Adobe Online Marketing Suite. Enterprise web applications tend to have different globalization requirements and constraints than consumer-oriented sites, and so when it comes to internationalization and localization, they must be approached in a somewhat different manner. Strategies that work well for consumer-oriented sites may not work well for enterprise web applications, and vice versa. This presentation examines the issues surrounding the internationalization and localization of large-scale, PHP-based web applications in an enterprise, SaaS context, using the Adobe Online Marketing Suite as a point of reference. The primary goal of the presentation is to show how enterprise internationalization and localization differs from non-enterprise internationalization and localization, and provide an overview of the different resources that are available to do enterprise internationalization and localization in PHP. By looking at some common internationalization and localization problems, and then describing how these problems were dealt with in the case of the Online Marketing Suite, I hope to demonstrate how PHP can be used in conjunction with other technologies like ICU to create fully internationalized, high performance web applications that are both easily localizable and highly scalable.

Presenter:

Track 2 - Localization Optimization Using Translation Repositories

Tak Takahashi
Globalization Engineering Manager, Teradata

Typically, localization requires the heavy involvement of translation companies to translate software resource strings as well as documentations. Translation of these strings and texts fully depend on TMX (translation memories) that are mainly utilized by translation companies. At Teradata, in order to optimize the localization process and minimize the external translation costs, we, a) pre-process translatable strings in various formats such as Java properties and .NET resx files, and then store them into a common format called the translation repository, b) "migrate" translations from older versions to newer versions automatically, c) "pre-translate" strings using translation repositories, and then d) submit to translation companies the strings with status flags to represent whether the string is new or modified for each resource string. Translation companies process and translate only strings that require new translation. This localization process and its supporting tools usually reduce the external translation costs for software by an average of 20-30 %. The process and tools were internally developed and have been since used for software localization at Teradata for years. We are now capable of handling DITA XML documentations in the same localization process as well.

Below is a list of challenges we face in software localization and translation:

- For frequently updated/released applications, we need to "migrate" translation from older versions to newer versions, for example from Teradata CRM 6.1.1.2 to 6.1.1.5. This migration process may or may not require new translation by the translation companies, but the language pack may need to be specific for that particular release. We want to automate this translation migration process.

- Cost reduction is always a challenge for everyone. We want to reduce the external translation costs we pay to the translation companies. When we translate software or documentation, every word sent to the translation companies will cost us, even if it was translated before.
- Once we have localized, for example, a product-A, we would like to share or reuse translations for other products, without involving the translation companies.
- We would also like to improve the localizability test and avoid any functional problems caused by inappropriate translation or localization.

To address these issues, we have designed and implemented a new localization process using the translation repository and its supporting tools. The translation repository was designed as a relational database that consists of tables, columns, and rows, so that we can scan tables and submit a query by SQL to search English or translated strings.

At Teradata, we now generate and maintain translation repositories for any software localization. The translation repositories enable:

- Localization migration
- In-house translation sharing
  - Reuse and sharing of translation
  - Facilitate in-house translation
- Translation using common resource file format for any software translation
- Facilitate validation and corrections of translations
- Localization problem investigation
- Reduction of external translation costs

In 2010, we have also enhanced our tools/process to handle DITA documentations. Now, we are capable of pre-processing DITA XML documentations, pre-translating elements using translation repositories generated from the software, and reducing the external translation costs.

Presenter:

Track 3 - Bidi Parentheses Algorithm

Murray Sargent III
Partner Software Design Engineer
Microsoft

Ayman Aldahleh, Gilead Almosnino, Peter Constable, Dylan Deverill, Andrew Glass, Michael Kaplan, Laurentiu Iancu, Dwayne Robinson, Murray Sargent, Robert Steen

Microsoft Corporation

The Unicode Bidi Algorithm (UBA) can display parenthesized text in strange ways, such as a)b) instead of (a)b, as described in a recent blog post*. That post describes an algorithm to fix the display of many such cases and that algorithm shipped with Microsoft Excel 2007/2010. The problematic cases have different UBA directionalities for the two parentheses of a matched pair. The algorithm essentially says that in such ambiguous cases, use the paragraph (or embedding) directionality for both parentheses. In all cases, increment the levels of text runs inside by 2 when necessary to keep the text inside the parentheses. In this presentation we describe an enhancement of the algorithm to deal with cases in which the text inside the parentheses has a single directionality. For such cases, the directionality of the parentheses is chosen to be the same as the directionality of the text within them. This refinement handles a set of anomalous cases where, for example, parenthesized English text appears in right-to-left paragraphs.

This improved algorithm displays the vast majority of parenthesized text the way one would want, but it is fair to say that no simple algorithm can handle all cases. The UBA has the LRE, RLE, PDF, LRO, RLO Bidi control characters to force particular choices. In the event that any of these Unicode Bidi control characters are used, the Bidi parentheses algorithm is not used, because the assumption is made that the user has specific choices in mind. The simpler Unicode LRM and RLM control

characters can be used with the algorithm if desired.
* http://blogs.msdn.com/b/murrays/archive/2010/05/07/bidi-paragraph-with-parenthesized-text.aspx

12:30-13:30 - LUNCH

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

Presenter:

Matt Sanford
Tech Manager of International Team, TwitterInc.

Track 1 - Internationalizing Twitter

Social networks connect you to people you already know. Twitter's model of connecting people to what's most important to them at any given time has created a cross-language and cross-cultural network unlike any other. All of this creates changes to language and communication that present a unique localization challenge. With our continued expansion in Japan, our translation of twitter.com, and our work on supporting Tweets in any language, we've learned some valuable lessons we're excited to talk about. Twitter should work in any language, even if we don't have the resources to support a full twitter.com localization, and making that happen is a large undertaking.

Presenter:

Shaopeng Jia
Senior Software Engineer Google Switzerland GmbH

Track 2 - libphonenumber - The Swiss Army Knife of International Telephone Number Handling

The libphonenumber project is an opensource project from Google, which provides Java, C++ and JavaScript APIs that supports parsing, validating and formatting international phone numbers for over 200 countries. This presentation will walk you through some common challenges when handling international phone numbers, and discuss hands-on how these challenges could be addressed with the APIs provide by the library. It will also present some common misconceptions with international phone numbers, and provides recommendations on best practices of handling phone numbers.

This is a 201 level talk compared to the talk I gave last year. It provides a more in-depth look at international phone number handling, and places focus on the new development of the library in the past year.

Presenter:

Roozbeh Pournader
Internationalization Engineer, Google

Track 3 - Bidirectionalization and Localization

In the process of bidirectionalizing a software application, the main challenge appears to be mirroring the interface properly and supporting all the nuances of the very sensitive markets. Getting the localized strings displayed properly usually gets overlooked. But it's a long and error-prone path, from translating a string by a linguist not familiar with the details of the Unicode Bidirectional Algorithm (UBA) to the application displaying the string to the final user. This talk will suggest best practices for localizing for the bidirectional markets, and cutting the cost caused by the expensive endless loop of finding bidi bugs in translated strings late in the QA process. A comparison of some existing tools and platforms will also be provided, together with suggested solutions to harder-to-handle issues caused by the UBA.

14:30-15:20 **SESSION 4**

Presenter:

Norbert Runge
Test Engineer, Google, Inc.

Track 1 - Encoding Health of the WWW

Web pages are written in many languages, using dozens of character encodings. In recent years the World Wide Web has steadily migrated towards the Unicode (UTF-8). This presentation examines the "encoding health" of publicly accessible web pages: What type of encoding errors are typical, what are the symptoms and

causes, how frequently do they occur, how should they be fixed? I will show techniques for finding and evaluating such problems, show how a search engine can work around them, and how they are exposed to webmasters.

For example, some pages are double-converted to UTF-8, turning 'ü' into 'Ã¼'. Many pages have the wrong encoding declared, others contain a mix of encodings. I will show details for this and other examples.

Presenter:

Track 2 - Internationalization in Google+

Mark Davis
Staff Test Engineer & other titles, Google, Inc

Luke Swartz
Product Manager, Google

Google+ launched from day one in over 40 languages. It adds a number of new internationalization capabilities, which we'll review in this presentation.

Presenter:

Track 3 - DecoType font concepts and design tools for Arabic typefaces

Thomas Milo
DecoType

DecoType are the first to develop novel, highly automated template-based design tools for Arabic fonts that cut development time to a minimum and eliminate the need for complex table building. By exploiting DT's unique smart font architecture, particularly porting a conventional font to the DecoType format can be done in a few hours. The result is a contextually fully programmed typeface that supports the complete Arabic block of the latest Unicode Standard. In addition to that, there is an advanced template that guides the designer to add the essential, but because of complexity omitted dissimilation features into her or his typeface as well. A dozen such typefaces have already been made for WinSoft Tasmeem.

15:20-16:00 - Afternoon Refreshments in Exhibit Area

SESSION 5

Presenter:

Track 1 - Web App i18n/l10n

Luke Swartz
Product Manager, Google

Web Applications--including those written for various mobile platforms--continue to grow in number and sophistication, in many cases displacing traditional desktop software. However, traditional tools and methods for making software work internationally do not always work for web applications. At Google, we are trying to solve the following problems:

- Core i18n Libraries: Giving web applications access to internationalization libraries on par with those available to desktop applications.
- Resources: Allowing web applications to easily package and manage strings and other international resources.
- Localization: Making it easy to translate web applications' strings and other resources.

In this talk, we will explore each of these challenges, and show some potential solutions, which will help web applications become fully international.

Presenter:

Track 2 - Localization Data Standards: Apathy, Skepticism, and Cynicism

Helena S Chapman
Program Director,

According to a 2009 European Union study, the language industry's annual compounded growth rate was estimated at 10% minimum over the next few years, resulting in approximate value of 16.5 billion to 20 billion € in 2015. For an industry

IBM

with such potential, it is difficult to comprehend the lack of technical leadership and investment in interoperability of localization data. In this session, we will explore why open standards and consistent implementation of these standards are important to your organizations, the current status of the industry support with regards to localization data interchange standards, and what can be done to reinvigorate the focus in this area.

We will also take time to examine where the data interoperability gaps are in an end-to-end content/data life cycle of a localization request. What roles can Unicode Consortium play in helping mature and drive the most appropriate adoption of its existing and future standards and assets. Most importantly, as a Unicode standard supporter, what you need to be aware in contributing to drive these standards to benefit the localization operation of your organizations directly or for your clients.

Presenter:

Track 3 - Unicode and the Revolution

Adil Allawi
Technical Director, Diwan Software Limited

What is now known as "the Arab Spring" has pushed a new wave of Arabic language users into the world of social media. But support for bi-di languages (like Arabic) on the web is difficult and can sometimes feels like a second class experience. What problems do these users face and what can the various web companies do to help them? The presentation will try to answer these problems and find a way forward for the future.

For the past five years I have been the Iraq correspondent to Global Voices and had the chance to work closely with those on the leading edge of the Arabic social media revolution. In this presentation I will seek to tie this experience with the technology that underlies it.

I will review the history of the Arabic language on the web, its current problems and successes and make a comparison between the uses of social networking sites in Arabic and their equivalent uses in mono-directional languages.For the presentation, I also hope to include feedback from active users of social networks on how they use this medium, what are their frustrations and successes on their chosen devices.

Finally I will conclude with a review of how Global Voices reports on social networks in multiple languages and make recommendations on how social web sites can improve Arabic support for their users and those reporting the conversations.

17:00-17:50 SESSION 6

Presenters:

Track 1 - Breaking The Language Barrier On The Social Web

Andrew Swerdlow

Co- Author:
Nav Jagpal
Technical Program Manager, Google

The social web has arrived with Social Network Sites (SNSs) such as Facebook having over half a billion users spending 700 billion minutes per month on Facebook. SNSs are no longer just for English users with recent reports of Twitter having more then 50% of Tweets in non-English languages. As social networks increase the diversity of languages published on their sites it can be difficult and confusing for users to understand each other. Social Translate is an open source project developed as a Chrome extension which attempts to automatically translate event streams and friends comments on SNSs. The extension allows users to select their primary language, when a user visits a social network site such as Facebook and Twitter it will use Google translate to detect the language of the event stream and and translate the text to the user's primary language. Social Translate also allow users to select multiple languages as their secondary language that will not be translated. This is useful for users that speak multiple languages and would like to have their events streams displayed in several languages. The Social Translate project serves as an interesting research case on how to combine machine translation (MT) to the real time social web in an open accessible way. This presentation will provide an analysis on the quality of machine translation on social networking sites such as

Twitter and provide the impact of MT on social media.

Presenter:

Michael S. Kaplan
Program Manager, Microsoft Corporation

Track 2 - Korean Hangul: from Sejong the Great's Hunmin Jeongeum to Unicode 6.1

Hangul has had a long history from the 1446 document that first described the underlying Jamo to the latest Jamo additions to the Unicode Standard. This presentation will do a whirlwind and only mildly irreverent tour of that history in the form of a presentation to Sejong to explain what has happened, highlighting the use, encoding, and re-encoding of one of the more perfect alphabets, imperfectly handled, in this or any other age.

Presenters:

Behnam Esfahbod
Yahya Tabesh
Sharif University of Technology

Track 3 - An Abstract Model for the Typography of Perso-Arabic Script

Perso-Arabic script, the second-most used writing system in the world, has many unique properties which have made its computation harder than some other scripts. In this paper we introduce an abstract model for the typography of Perso-Arabic script which exhibits the hidden properties of the script and makes the typographical computation of Perso-Arabic text possible. This properties have been ignored in most of the recent works on Perso-Arabic script, specially in the international standards.

Each letter in Perso-Arabic script, in any of its cursive forms, is constructed from a Base Shape and some of the them are accompanied with some Auxiliary Shapes (for example: Dots, super-script and sub-script Alefs, Madda, Hamza). These Auxiliary Shapes may appear on the top of the Base Shape, on the bottom, and sometimes on the head or the tail. Also, some other Auxiliary Shapes may accompany each letter as a separate Unicode character (for example: Harakats, Shadda, Sukun) which also will be positioned above or below of the Base Shape.

Base Shapes and Auxiliary Shapes work as the building blocks of the Perso-Arabic script. These visual properties are consistent in all of the writing styles of Perso-Arabic script (like Naskh, Thulth, Nasta'liq, and Tahriri). In this Abstract Model, we have encoded these properties such that a series of Shapes can be computed for any Unicode string.

Also, we introduce a metric distance based on our Abstract Model, the Shape Distance, for strings of Unicode characters. This metric makes it possible to compare Perso-Arabic strings based on their actual appearance, regardless of what writing system or font is used. The Shape Distance works such that strings with very similar appearance would have a distance close to zero, and big difference in the appearance results to large numbers. For example, two words with similar letters which only differ in one auxiliary part (like a dot) would have less distance than strings with letters that have different Base Shapes.

And finally, we have studied two Persian text corpora based on the properties of the words, the letters, and the Base Shapes and Auxiliary Shapes. We show that the distribution of Base Shapes and Auxiliary Shapes follows the same the pattern as distribution of the letters.

The first important application for our model is the security of Perso-Arabic domain names. By the introduction of Internationalized Domain Names (IDNs) and looking at the future of the internationalized Internet, security of domain names (at both the TLD level and the registry level) has became a serious concern of ICANN and various ccTLD ang gTLD registries. The Abstract Model and the Shape Distant algorithms can be used to calculate the similarity of Perso-Arabic domain names with very high accuracy.

The second application for our model is the font industry, it can be used in font generation, alteration and verification programs. Using this model, some parts of the

glyph generation, classification, and table generation in Perso-Arabic fonts can be automated.

Another application of this model is the font rendering engines which at the moment depend on ArabicShaping table in UCD and the tables provided in font tables. The character data provided in ArabicShaping table is incomplete for some use cases, and the font tables cannot always be trusted to be complete or accurate. This model can help in better Perso-Arabic text rendering and increasing the stability of font rendering engines for this script.

Notes

1. "Typography" here means "the general character or appearance of printed matter."

2. Perso-Arabic script is called "Arabic" in Unicode standard. Also some other names has been introduced, like "Arabetics". In this paper we use "Perso-Arabic" as the name of the script to distinguish it from the Arabic language.

18:00-20:00 - IUC35 CONFERENCE RECEPTION (IN EXHIBIT AREA)

Wednesday, October 19, 2011

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

Presenter:

Andrew Glass
Program Manager, Microsoft Corporation

Track 1 - International User Experiences in Windows

We have entered an era of explosive growth in software usage among speakers of languages other than English or other major European languages. In addition, a large segment of that growth includes users with dual- or multiple-language needs. In this context, it is increasingly important for an operating system to support great language-related experiences for worldwide and especially multi-lingual users. In this talk, we will outline some challenging areas of language support centering around text input and reading experiences, and introduce some possible solutions in these areas being explored for the Microsoft Windows platform.

Presenter:

Su Liu
IBM

Track 2 - Study for Processing Unicode Data with Multiple Versions of Unicode and Non-Unicode Standard

Unicode encoding and code set conversion are key features in solutions of storage, information retrieval and data mining systems. To process and support Unicode data with multiple versions of Unicode and Non-Unicode standards are challenge tasks in storage (e.g. digit libraries and data centers etc…), which contains the data in current version and at least one earlier version. Meanwhile, some Unicode side effects, such as variant characters, PUA, and overheads on data normalization and conversion, aggravate complexity to solve the multiple version issue. This paper discusses the multiple version impacts and Unicode data processing strategies on levels of storage, network and OS.

Keywords: Unicode Data, Multiple Versions, Network, Conversion

---

Presenters:

Pim Rietbroek

Track 3 - Fonts for the Ages

A scholarly publisher needs to be able to publish any text in any European

John Hudson
Brill / Tiro Typeworks

language from any period, and to do so within traditional canons of typographic quality and sophistication, as expected by their authors and readers. This presentation describes the demands this makes on the publisher, and on editors and authors, and looks at how these demands are met through careful specification of requirements, standardisation on Unicode text encoding, and development of extensive and typographically sophisticated OpenType fonts.

Pim Rietbroek (Brill) and John Hudson (Tiro Typeworks) present an overview and selection of short case studies from their five-year project, illustrating some of the challenges encountered in texts ranging from transliterated ancient Egyptian and Sumero-Akkadian to linguistic descriptions of present-day endangered languages, with stops among ancient Greek acrophonic numerals and the not-always-helpful Unicode encoding unifications and disunifications.

10:00-10:50 SESSION 8

Presenter:

Michael S. Kaplan
Program Manager, Microsoft

Track 1 - Locales on Windows - the view from 18 years in

It was 1993 that the basic model for locales was integrated into Windows in its current form, and that model has been largely unchanged for much of that time. In this unique view of those 18 years, you can find about about the lessons learned, unlearned, relearned, and mis-learned. You'll leave this all up view feeling both more impressed and more embarrassed to know Microsoft than you ever have before, even if you were there while it was going on!

Presenters:

Leandro Reis
Senior Program Manager, Globalization. Adobe Systems

Mike McKenna
Senior International Engineering Manager, Zynga

Paul-Henri Arnaud
Senior Process Analyst, Autodesk

Track 2 - Internationalization assessments: Merging the best of three approaches

Software globalization seems to many to be either a black-art practiced by an esoteric guild of polyglot bit-twiddlers or something you get for free because you happen to use Java and Unicode. The truth is neither, and it is something that can be implemented in a methodical way that can be measured. Three different corporations - Adobe, Autodesk and Zynga - are tackling the problem of how to measure globalization compliance and progress across a wide range of technologies and products. They decided to join together in an open-source fashion to decide on a standardized set of requirements, mappings to specific technology genres and method of grading for software globalization. This presentation will present their efforts to date with discussion on similarities and differences among approaches by the companies, as well as issues encountered, solutions implemented and solutions envisioned.

Presenter:

Peter Constable
Senior Program Manager, Microsoft Corp.

Track 3 - Creating World-Ready Apps For Windows 8

The next version of Microsoft Windows is in development and promises to provide great new opportunities for the Windows developer ecosystem. In addition to introducing new app development paradigms, Windows 8 also adds a lot of new multilingual and globalization functionality. This talk will provide an overview of additional functionality that developers can leverage to develop world-ready desktop or Metro-style apps for Windows.

10:50-11:10 - Morning Refreshments

11:10-12:00 SESSION 9

Presenter:

Peter Constable
Senior Program

Track 1 - Application Resources and Localization for Metro-Style Apps in Windows 8

Application developers still face big challenges in creating and deploying localized,

Manager, Microsoft Corp.

multilingual apps. These challenges have been one of the focal points as we have worked on the next version of Microsoft Windows, Windows 8. This talk will take a detailed look at the new application resource model in Windows 8 for Metro-style apps created using HTML or XAML. You'll leave with a basic understanding of the new resource infrastructure and how it makes localizing your app a lot easier.

Presenter:

Steven R. Loomis
Software Engineer, IBM

Track 2 - What's new in CLDR 2.0

The Common Locale Data Repository is a project for the exchange of language and locale information used in application development, and to gather, store, and make such data publicly available. By pooling resources, the time and expense of collecting good data is minimized, and language groups have an avenue to get their data into implementations. This session will discuss implementation of CLDR and the latest project status, and how the process is being improved to produce higher-quality data. Panelists will then discuss how they are making use of CLDR data, the latest project status, and issues in the collection and production of data. The panel will consist of persons from multiple vendors involved in deploying CLDR in their own products and projects, as well as those involved in the data gathering and vetting process. Comments and questions will be welcomed from the audience.

Presenters:

Adil Allawi
Technical Director,Diwan Software Limited

Track 3 - Developing a Unicode font for the desktop, the mobile and the web

The only Arabic font I ever designed, Geeza, started life in 1985 as one of the standard fonts on the original Apple Arabic Macintosh and is still the standard Arabic font on Mac OS and iOS. This presentation covers my approach to developing Geeza over the years, automating the font creation with Unicode data. I will also cover my work with Typekit.com to publishing Geeza as a multilingual web font. The talk will be aimed at giving information generally relevant to developing multilingual, Unicode fonts on different media and platforms.

Over the following 25 years the font has been rebuilt several times over, extended to the full Arabic Unicode range, ported to every kind of device. The latest incarnation is now a web font for every major browser. The design needed to be relevant for user interfaces, printing and small screens yet still appear compatible with Roman fonts. Support for the full Unicode Arabic range, meant adding an extra 1500 glyphs together with the relevant tables for Arabic shaping, ligatures, justification and kerning.

Along the way I will explain the features need for a modern typeface to be useful in a world where data may be exchanged across different standards; How to approach the development of user interface fonts and the new standards for web fonts. I will discuss the importance of embedding semantic information into a font to allow unique identification of its glyphs and allow equivalence to be found in other fonts.

The presentation will conclude with a discussion about the future for international typefaces as they make their way into open standards and become part of the content of the worldwide web.

12:00-13:00 - LUNCH

SESSION 10

Presenter:

Doug Felt
Google

Track 1 - Best Practices with the Java 7 Locale

In Java, the Locale class is fundamental for developing global software. Java7 adds several important enhancements to resolve issues that were difficult to handle in previous versions of Java. For example, the new script field allows

Steven Loomis
Software Engineer
IBM

.

developers to package Chinese localized resources in logical manner, and full support for BCP 47 language tag conversion allows software to exchange language and locale information through standard protocols without any data loss. This session provides a brief overview of the enhancements, followed by best practices and programming tips recommended for Java application developers.

Moderator:

Markus Scherer
Unicode Software Engineer, Google Inc.

Mark Davis
Sr. Internationalization Architect, Google Inc.

Track 2 - Plural & Gender in Translated Messages

"There are 1 file(s)." / "Alice added 1 people to his mailing list." - User-facing messages with placeholders for numbers and strings are common technology. These require the placeholders and text to be reorderable to account for grammar of different languages. However, the common technology does not solve the problem of plural and personal gender in placeholders. That is, depending on the language and the placeholder values, the surrounding text often needs to change, as illustrated by the examples above.

ICU has been improving on the Java formatting framework, adding support for such message variants in both its C++ and Java versions. In addition, other aspects of message formatting have been simplified. This session explains the challenges, approaches, and new functions and capabilities.

Co-authors: Markus Scherer & Mark Davis

Presenter:

Anshuman Pandey
Ph.D. CandidateUniversity of Michigan

Track 3 - A Pre-script-ion for the Future: Unicode and the Development of Minority Languages in South Asia

South Asia is home to tremendous linguistic diversity. 'Ethnologue' records roughly 438 languages spoken in India alone. According to UNESCO's Endangered Languages Programme, 198 of those languages are considered to be endangered and there an additional 140 languages with similar status in the region stretching from Afghanistan across to Nepal and Bhutan. The conditions of these languages and the ongoing decline of other minority languages, which are not yet endangered, are the partial result of inadequate institutional support and development. The growth of new digital technologies has the potential to positively alter the course of endangered and minority language by offering speakers innovative ways in which to use and maintain their mother tongues. The basis of these technologies is Unicode. The aim of the Unicode Standard is to encode characters that are needed for representing text in all modern writing systems, as well as most historic scripts (The Unicode Standard, Version 6.0, p.10). Unicode, then, is a prescription for a brighter future for minority languages in South Asia

This presentation will discuss the role of Unicode in the development of minority and endangered languages in South Asia, with particular focus on India. The talk will begin with an update on the support for South Asian scripts in Unicode 6.0 and an analysis of how well the Standard covers these scripts today. It will then present five current character-encoding projects for writing systems used by minority language communities in India and Nepal and discuss the conditions of these languages, the linguistic requirements of these communities, and the potential that Unicode offers for increasing education and literacy using the language. The talk will then describe the role of Unicode in the Government of India's National e-Governance Plan (NeGP) and how the plan has the potential to facilitate institutional and governmental support for endangered and minority languages. The talk will close with a presentation of several case studies of writing systems being created today in India by minority language communities and what such ongoing activity will mean for Unicode. The presentation will conclude by discussing the current projects of the University of California - Berkeley's Script Encoding Initiative, which is working with the user communities

to propose these scripts for encoding into Unicode.

14:00-14:50	SESSION 11

Presenter:

Track 1 - Agile Internationalization and Localization

Michael Kuperstein
Localization Engineer, Intel

Agile development methodologies are swiftly being adopted throughout the software development industry. This presentation will illustrate key concepts, challenges, and solutions for performing Agile internationalization and localization. One important goal of Agile development is to have release-ready software at the end of each 'sprint', which typically have a duration of between one and four weeks. For example, if a feature involves building a contacts list, then entering, storing and sorting of contact names should work for all languages, even for a single-language product. If the product is also being localized, then we would expect each particular feature to be fully localized into all the target market languages at the end of the sprint. In other words, "Done" truly means "No work left to be done." Unfortunately, this puts even more pressure on internationalization and localization teams, since the scheduling and quality challenges escalate rapidly as sprints become shorter and language counts climb. We'll cut through the confusion to focus on a handful of proven internationalization and localization strategies that can ensure a great user experience for customers of every culture.

Presenter:

Track 2 - Speech Internationalization at Google

Martin Jansche
Staff Software Engineer, Google Inc.

Speech Internationalization at Google

Pedro J. Moreno, Linne Ha, and Martin Jansche*
Google, Inc. (*corresponding author)

Internationalization (i18n) and localization of software that processes spoken input or output is faced with challenges that differ from those found in many other software i18n projects. In this talk we describe our experience with internationalizing Google Voice Search, a speech-to-search service available on many popular mobile devices. First launched for US English in November 2008, Google Voice Search is currently available in more than 20 languages.

Our lack of precision in the number of available languages illustrates one the main challenges of this project: working definitions of "language" and "dialect" vary depending on context. Focusing purely on technical aspects, our notion of language is driven by the current limitations of speech recognition technology. For example, Google has developed speech recognition models for several variants of spoken English, including American, British, Indian, Australian, and South African English. But our support for English does not end here. Many other languages have been influenced by English: Our Cantonese recognizer was explicitly designed to deal with the many English loanwords that are in daily use in Hong Kong. The languages of Europe all borrow English words liberally, despite the efforts of local language academies.

This kind of linguistic diversity makes the challenge we face even harder, since it increases the number of language projects we have to deal with. This is compounded by the fact that internationalizing speech technology is hard because bringing up a recognizer in each new language is a separate development effort, requiring significant amounts of data, compute power, and engineering time. We are fortunate to work in an environment where enormous compute power is available and can be easily harnessed. Of the other two factors, we will assume that engineering time is always limited and at best, we can reduce the accidental complexity of the very complex task of building recognition models. Since our development approach is heavily data-driven, most of our recent progress has been in the area of data acquisition, where significant changes have been made

to the tools and processes which we use to acquire linguistic data as needed for building recognizers.

We generally need three kinds of data: spoken utterances plus textual transcriptions; pronunciation dictionaries; and large amounts of text. We have crowd-sourced several aspects of our data collection efforts. For example, we generally need spoken examples for a given target language, from a variety of speakers, and under varying environmental conditions. We have built tools that make it easy to collect such data from volunteers. We have also made process changes to allow us to manage the data collections remotely while monitoring progress and assessing data quality. This has allowed us to collect acoustic data many times faster than before.

The collection of pronunciation information cannot be crowd-sourced as easily since the current process requires a certain amount of linguistic expertise. Volunteers are asked to transcribe words into crude phonetic representations. For some languages, e.g. Spanish, the pronunciation of a words is readily apparent from its orthography and can be expressed algorithmically. In those cases we use ICU transforms to give us word pronunciations, perhaps combined with a brief pronunciation dictionaries of exceptional or foreign words.

While text resources are often easily available, this does not hold for all languages. For example, Voice Search is currently available in South Africa in English, Afrikaans, and Zulu. For Zulu, the amount of text available in electronic form is considerably more limited than for the other languages. We'll share our thoughts on what could be done to encourage the creation and dissemination of data for under-resourced languages.

Presenter:

Martin Raymond
Script Information Engineer/Editor, SIL International

Track 3 - ScriptSource: Making information on the world's scripts accessible

Although there is plenty of script information on the web, there has been a need for a web site to present the information authoritatively and clearly, making it easier to understand the often complex relationships between scripts, characters and languages. ScriptSource has been designed to meet that need and to answer questions such as: 'Which scripts can be used to write that language?', or, 'Which writing systems use this Unicode character?'. The site allows registered users to add information to the site in the form of entries, which may include links to other sites, all entries being moderated. Users can also post 'needs' to enlist help in solving script-related problems. ScriptSource imports language data from the Ethnologue, character data from Unicode and locale data from the CLDR (Common Locale Data Repository). CLDR's locale data, such as exemplar sets, is linked to the scripts, characters and languages it relates to.

This session will cover some of the needs ScriptSource has been designed to meet, as well as the challenges encountered in bringing together information from different sources and creating the data associations to make it as meaningful as possible. There will be a demonstration to show how easy the ScriptSource User Interface is to navigate and to illustrate the main functions, including adding information about a script. The use ScriptSource makes of CLDR data will also be demonstrated, and the plans for a more extensive interface with the CLDR will be discussed.

14:50 – 15:10 - Afternoon Refreshments

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

Co-Presenters:

Mark Davis

Track 1 - Bits of Unicode: ICU Data, Algorithms, and Performance

Supporting Unicode with good performance and with reasonable memory footprint

Markus Scherer
Google

presents a challenge. No matter which encoding form is used to represent Unicode, 1,114,111 different codepoints and associated data are a lot to handle. Most classic data structures are byte oriented, which is often not optimal for dealing with Unicode, even with UTF-8.

ICU uses a number of innovative algorithms and data structures to handle internationalization, balancing tradeoffs between performance and data footprint. This presentation covers some of the more interesting of these structures, and their applicability beyond internationalization. The discussion includes: the new trie structure for string lookup in ICU, effective use of inversion lists and inversion maps, compact character mapping tables, transliteration mappings, and others. Code that supplies and uses these structures is part of ICU, the Java/C/C++ open-source Unicode enablement library.

Co-authors: Markus Scherer & Mark Davis

Presenter:

Ken Lunde
Senion Computer Scientist, Adobe Systems Incorporated

Track 2 - Genuine Han Unification

There have been major shifts and reforms in East Asian writing systems in the past that seemed revolutionary at the time, but that are now considered to be standard and thus completely acceptable. One such reform was the hanzi simplification in China that took place during the early 1950s. Thanks to the Web and other advances in communication technology, the world has become a smaller place. Thus, more cross-cultural interaction is taking place than ever before. Perhaps serving somewhat as a catalyst, Unicode, with its tens of thousands of CJK Unified Ideographs that cover the needs of virtually all customers of the locales that use them, provides the foundation for another shift or reform, though it is not likely to take place for another decade or two.

Today, one can easily argue that for a single font to adequately serve multiple CJK locales, it must include more than one glyph per CJK Unified Ideograph code point. Such fonts are referred to as Pan-CJK fonts, because they serve the needs of more than one CJK locale. I have predicted that at some point, years or decades into the future, cross-cultural interaction will evolve into initiatives whose aim is to genuinely unify CJK Unified Ideographs across all CJK locales. This is likely to have the effect of making a single glyph acceptable for all CJK Unified Ideographs. The Chinese standard designated GB 18030 is actually a step in this direction, mainly because it specifies a single glyph for each CJK Unified Ideograph code point.

This presentation will explore the history and development of ideographs and Han Unification, and draw conclusions based on the presenter's own experience developing CJK fonts and working with CJK character set standards.

Presenter:

Roy Boney, Jr.
Language Technologist, Cherokee Nation

Joseph Erb
Educational digital media specialist, Language Technology Program at Cherokee Nation Education Services Group

Track 3 - The Cherokee Syllabary in Digital Applications

With the inclusion of the Cherokee syllabary in Unicode, it is being in used in some of the most popular and advanced devices in the world such as the iPhone and iPad. The Cherokee syllabary was included as part of the Unicode Common Locale Data Repository version 1.8. This has helped in the perpetuation of an endangered language and is paving the way for a renaissance of the Cherokee language in modern digital media culture.

One problem of the digital globalization of communication is that it has the potential to erode the already endangered cultures and languages of indigenous peoples. These communications rarely occur in the indigenous languages of minority cultures, if at all. It does not need to be this way, and proper adoption of the technology by the community is paramount for success.

Jeff Edwards
Language Technologist, Cherokee Nation

This presentation will discuss the efforts undertaken by the Cherokee community to address these problems. It will discuss the adoption of the Cherokee syllabary into Unicode, the usage of the Cherokee syllabary in modern computing systems, the mobility of the Cherokee language in various hardware platforms and social web presences, and the adoption of Cherokee language technology by the Cherokee community.

The presentation will be by Roy Boney, Jr. and Joseph Erb, Language Technologists of Cherokee Nation Education Services Group.

16:10 - 17:00 SESSION 13

Presenter:

Stuart Gill
Member of Technical Staff, Google Inc.

Peter Edberg
Senior Software Engineer, Apple Inc.

Markus Scherer
Unicode Software Engineer, Google Inc

Track 1 - New in ICU

The International Components for Unicode library, or ICU, provides a full range of services for Unicode enablement, and is the globalization foundation used by many software packages and operating systems, from mobile phones like Android or iPhone all the way up to mainframes and cloud server farms. Freely available as open-source, it provides cross-platform C, C++, and Java APIs, with a thread-safe programming model.

This presentation will provide a brief overview of ICU, with emphasis on the recent updates in ICU 4.8, including the latest support for Unicode 6.0 and CLDR 2.0, collation reordering for better customization and reduced collation data, plural and gender in messages, and other changes (see http://icu-project.org/download/4.8.html). The presentation will also touch on ICU's planned direction for 5.0 and future releases.

Presenter:

Zhang Zhoucai
CEO, Beijing UniHan Digital Tech Co.Ltd

Zhang Chiyi
COO, Beijing UniHan Digital Technology Co., Ltd.

Track 2 - Some Special Requirements for Cloud-Publishing of Chinese Ancient Classics

Cloud computing focused in the early going on software as a service (SaaS) applications, but Amazon, Netflix, Google, Apple, Microsoft and others are now tapping the cloud for content delivery (some of these companies focus on streaming entertainment, while others focus on content creation/management). Both e-publishers and their readers increasingly rely on web, and want to get more and more benefits from so-called cloud-publishing. As a typical culture heritage, a variety of huge data base of Chinese ancient classics, such as ?????? and ??????? have been built up and providing web service in world-wide? They are regarded as earliest breaking through in this field, which were almost made immediately after the Unicode/CJK unification standard released. Now, this kind of data base are facing the challenge and mission of migration in the cloud for digital transition and broader consumptions.

As Unicode/CJK standard developer and Unicode implementer in Chinese e-publishing industry, the author summarized their experiences and lessons in Unicode based e-publishing, and points out that, besides the ordinary requirements for general content provider on the web-cloud, Chinese Ancient Classics have some special requirements for Cloud-Publishing for better reading and usage experiences, which include but are not limited to

- super-CJK Font on Web-Cloud and its alternative solution,
- super-CJK Hanzi Handwriting Recognizer on Web-Cloud ,
- a Dictionary of Simplified-Traditional-Variant Hanzi on the web, and especially,
- an on-the-web OLD TERM CHECKER would be very much desired in order to help reader to judge and understand strange Chinese term's attributes - a dynasty year, a name of an officer, a name of a place, or a name of a

person?

In addition, Author will introduce their implementation of web-cloud based handwriting recognizer (UniHan Q-Pen) and the above mentioned TERM CHECKER built in their UniHan Classics data base on the web.

Co-Author:
Zhang, Chiyi, COO,
Beijing UniHan Digital Technology Co., Ltd

Presenter:

Tex Texin
Chief Globalization Architect, Rearden Commerce, Inc.

Track 3 - Does it hurt when I do this? Data for I18n Testing

This presentation recommends specific data values that are likely to identify internationalization problems in software intended for global markets.

Based on years of global software experience, these data values are useful in functional or linguistic QA tests of internationalized software. The data value recommendations include character encoding, postal address, locale and other data types typically used in software and will assist in finding common internationalization problems. This presentation will offer specific test suggestions.

Program is subject to change.

- To Register for IUC35: http://www.unicodeconference.org/registration.htm
  Or, contact Suzanne Leon at suzanne@omg.org
- Exhibitor Information: http://www.unicodeconference.org/be-exhibitor.htm
  Or, contact Ken Berk at ken.berk@omg.org

- Sponsor Information: http://www.unicodeconference.org/be-sponsor.htm
  Or, Ken Berk at ken.berk@omg.org, or 781-444-0404.

[iuc35/_borders/bottom.htm]