home
conference program
hotel
registration
become a sponsor
become an exhibitor
press room
past conferences
send me more info

Hotel cut-off:
09/26/2010

Venue:

Hyatt Regency Hotel
Santa Clara, CA USA

Program - Session Descriptions

Monday, October 18, 2010

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:

Sumit Sarkar
i18n Product Specialist

Track 3: Internationalization in Database Drivers for Your Applications

Everything you want to know about i18n and database drivers across ODBC, JDBC, and ADO.NET. Discussion starts by asking what Unicode support encompasses at the Database Access API level, and what components affect Unicode Support. Take a closer look under the covers at the low level data access across major RDBMS including DB2, SQL Server, Oracle, and Sybase. This includes identifying who is doing the conversions at each component of the data access application layer. To summarize and apply the learned concepts, host will answer key questions about your globalized application's data access: Why should conversions be avoided when possible; and what high level features of a database driver are recommended?

10:30-10:45 - Morning Refreshments

Presenter:

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

Richard Ishida
Internationalization Lead, W3C

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:

Ram Viswanadha
Internationalization Architect Yahoo!

Track 3: What's that gobbledygook in the URL?

Globalization of websites is going mainstream. Some sites are targeting multiple languages and some are targeting specific regions. Website owners are trying to find ways to make their properties more discoverable among the thousands of websites. They are constantly grappling with questions such as:

Does the URL structure and format have affect on search engines?
What are different ways to hint search engines to index, classify and categorize content?
Should IDNs be used? Should code points outside of ASCII range be used?
What are right metrics for measuring success?
What are issues with the current technical solutions?

This paper answers these questions, analyzes the issues with URL design and SEO strategies for globalized sites and presents solutions that can improve the accessibility and discoverability of websites.

12:30-13:30 - LUNCH

13:30-15:30 AFTERNOON TUTORIALS

Presenters:

Craig Cummings

Mike McKenna
Internationalization Architects Yahoo! Inc.

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 will cover: 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 International Components for Unicode (ICU) 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 to support more minority scripts, languages, and help solve linguistic processing issues.

Presenter:

Tex Texin
Xen Master XenCraft

Track 2 - 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, 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

Track 3 - Building Multilingual Websites in Joomla [Drupal]

A practical look at the language and locale capabilities of Joomla! and Drupal, two leading free software content management systems (CMSs). They let you build more

Jim DeLaHunt & Associates

powerful, more international websites faster. We look at: their core services for internationalization and locale support; localization of UI and content; and localization support in some leading modules. 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, for web designers and developers, and for managers of international web site teams.

15:30-15:45 - Afternoon Refreshments

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

Presenters:

Craig Cummings
Mike McKenna
Internationalization Architects
Yahoo! Inc.

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

Presenter:

Craig Rublee
Sr. Globalization Architect
Adobe Systems, Inc.

Track 2 - Developing World Ready Applications for Smart Devices

This tutorial will show how to use the Adobe® Flash® Platform tools and technologies to create a world ready application that can run in the browser or as a stand-alone application on mobile devices and desktop platforms that support the Adobe Flash Player plug-in and Adobe AIR®. During the tutorial, participants will create a sample application that illustrates how to externalize strings and other resources into separate localizable files and make use of ActionScript® to perform language sensitive string comparison and locale specific date, time, and number formatting. The session will also include topics on auto-layout of UI elements, UI mirroring, font selection, and support for vertical writing and complex scripts.

Presenter:

John Emmons
Senior Software Engineer
IBM

Track 3 - ICU Workshop

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.

18:00-19:00 - Welcome Reception

Tuesday, October 19, 2010

09:00-09:15	***WELCOME & OPENING REMARKS***
09:15-10:00	***KEYNOTE Presentation***

10:00-20:00 - EXHIBIT AREA OPEN

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

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

Presenter:

John Yunker
President
Byte Level Research

Track 1 - Improving the Global Gateway: Established and Emerging Trends in Multilingual Navigation

Many web sites now support 30 or more languages and yet most web users speak only one language. To ensure that users can easily find (or change) their language settings, many companies rely on visual "global gateways" as well as IP and browser detection on the back end. Over the years, a number of informal best practices have emerged.

This session draws on more than seven years of research on global gateways, documenting established best practices. In addition, this session highlights emerging trends, and addresses questions such as:

- How can companies make their localized content more discoverable?
- When should companies avoid using geolocation?
- The challenge of displaying (and sorting) multi-script language lists
- Why do web sites continue to use flags?

Presenter:

Kirti Velankar
Yahoo Inc.

Track 2 - Internationalization with PHP

PHP is one of the most prominent and popular platforms for modern Web development. This updated session discusses PHP from the perspective of internationalization, what some of the challenges in PHP are, the features available in PHP 5.

This session also includes examples and usage in practical scenarios. You will learn how to effectively build applications for multiple languages and cultures using PHP with some of the internationalization features such as locales, sorting, resource bundles, as well as date, number and message formatting.

Presenter:

Ken Lunde
Senior Computer Scientist
Adobe

Track 3 - The Power of "Plain Text" & the Importance of Meaningful Content

Unicode has accomplished what no other character set or encoding can claim, in that it has been extraordinarily successful in accomplishing many of its goals, and it continues to develop to rise to new challenges. Regardless of how text is stylized using various applications or markup languages, the ability to search, copy, paste, import, and export is crucial for using and repurposing documents and their text. In other words, embracing Unicode's representation of "plain text" will ensure user-friendly digital data throughout a document's workflow and lifetime.

This presentation will cover the basics of "plain text" and its importance in providing meaningful content, and will explore some recent developments in Unicode that allow otherwise unencodable characters, such as variant forms of CJK Unified Ideographs, to be reliably represented in a "plain text" paradigm through the use of the Ideographic Variation Database (IVD). Examples from other scripts will also be provided. Various "plain text" pitfalls and bad-practices that undermine Unicode's success, such as PUA usage, CJK Compatibility Ideographs, and code point poaching, will be touched upon. Finally, emerging environments that thrive on "plain text" data, such as mobile, will be discussed.

11:30-12:20 SESSION 2

Track 1 - Global Ready Assessment Tool

Presenter:

Michael McKenna
Internationalization Architect, Yahoo! Inc.

A case study and discussion of developing a globalization process, tracking, and compliance system that is accepted by management, used by engineering, and helps drive products towards a goal of robust internationalization.

The globalization engineering team at Yahoo! was confronted with the daunting task of providing tools and technology to help bring hundreds of legacy products up to date with respect to international web standards and global customer needs.

Globalization Engineering created a Master List, and applied it across each step of the Product Life Cycle. They built a tool that through an interactive interview process presents a score card with suggestions on how to increase the globalization score of each product.

This talk will discuss the design decisions in the creation of this tool, as well as shareable contents of the master list.

Presenter:

Nebojša Ciric
Software Engineer, Google, Inc.

Jungshik Shin
Software Engineer, Google, Inc.

Track 2 - Making Javascript Multilingual

ECMAScript is a crucial component in websites and web applications. Unfortunately, the internationalization support in ECMAScript is too weak to meet the needs of modern programs. For example, there is no support for linguistically-correct sorting, multiple locales, all the needed date or time formats (such as month+day), and so on. This lack of support forces developers into expensive workarounds, involving either substantial code and data downloads, or accessing servers for needed operations. Both of these cost in terms of speed, memory, and complexity.

This is especially frustrating because all the major browsers run on fully capable client OSs, or have built in support like ICU. So the functionality is typically there, but there is no way to access it.

Our goal is to develop better client-side i18n APIs for ECMAScript, and eliminate the need for big ECMAScript libraries or server-side processing. Eventually, we want to make them a part of WebAPI (worked on by WebApp WG) or ECMAScript standard.

This presentation covers the current status of ECMAScript I18N support, and describes and demos our approach to dealing with the problems, and indicates what the future could hold.

Presenter:

Rusmin Dirgantoro
Chief Engineer
Aeontera, Inc.

Track 3 - Multilingual Unstructured Text Analytics

Textual unstructured data is corporate information that lives in email messages, documents, news articles, blog and forum posts, technical support case logs, instant message scripts, etc. Survey says that 80 percent of business-relevant information originates in unstructured form. This valuable information can be analyzed to improve customer relationship management and to better understand customer wants, needs, and dislikes. Unstructured text analytics is a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, investigation, and predictive intelligence.

According to "IDC Enterprise Disk Storage Consumption Model" report released in Fall 2009, while structured data is projected to grow at a compound annual growth rate (CAGR) of 21.8%, it is far outpaced by a 61.7% CAGR for unstructured one. The sheer amount of data increases the effectiveness of statistical-based analytics. However, named entity recognition, cross-language pattern search, transliteration, and syllable boundary detection, to name a few, remain difficult problems to solve. The speaker will present the challenges and solutions in unstructured text analytics in general and then focus on the multilingual aspects.

12:30-13:30 - LUNCH

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

Presenters:

Andy Bell
welocalize

Track 1 - Global by Design: Systematic Approach to I18N Testing

Abstract coming soon!

Presenter:

Mihai Nita
Globalization Architect
Adobe Systems, Inc.

Track 2 - New Internationalization Functionality in the Adobe Flash Platform

This presentation will cover the ActionScript classes in the new flash.globalization package. These classes allow for locale-sensitive number/currency/date/time formatting, sorting and case conversion using the operating system services.

We will demonstrate most of the functionality of the classes, and discuss implementation, with pros and cons, results, trade-offs, issues, and some of the surprises we encountered.

We will also touch on globalization functionality in Flex SDK version 4, UI mirroring, and the Flash text engine.

Presenter:

Erkki I. Kolehmainen
Kotoistus and Oy KREST Sales and Consulting Services

Track 3 - Name Writing in Public Registers for Multi-country Use

Personal names should be spelled correctly for both politeness and legal reasons. Unicode provides the technical capability, but the legal aspects bring other factors into consideration.

Within the European Union's internal market, people can move freely to any country,

Ltd

where they will eventually be registered in order to receive the public services that they are entitled to. For eGovernment services, EU-wide interoperability of the registers is a real requirement. Names that contain letters that are not part of the local alphabet, however, can cause considerable problems for their use as such. Some foreign letters, e.g., ß, þ, ð, ?, and ?, may actually be practically unrecognizable by the civil servants in several countries.

For this reason, a European standardization project is being put in place to define a common core repertoire and the full Latin repertoire for use in the public registers within EU. The project will also define the fallback transformation principles for those characters of the full repertoire that the Individual states may decide not to include in their national registers. For those nations that use non-Latin script, currently Bulgaria (Cyrillic) and Greece (Greek), the EU-wide interoperability will be based on the romanized forms used in their official travel documents. The same principle would apply to names of persons originating from outside EU.

The presentation will discuss the current line of thinking for the rationale for inclusion in the common core as well as some of the transformation principles for use with the rest of the full repertoire.

Although not necessarily of equal importance, the same solution would also meet the requirements for the correct spelling of the names of organizations and companies and their products as well as postal addresses.

14:30-15:20 SESSION 4

Track 1 - Global by Design: Systematic Approach to I18N Testing

Presenters:

Loïc Dufresne de Virel
Localization Strategist, Intel Corporation

Michael Manca
Localization Program Manager, Intel Corporation

Once the request to internationalize an application makes it into the Architecture Specifications, the Product Requirements Document, or the Low-Level Design Specifications, developers sometimes struggle to find their way through the labyrinth of issues they need to address; however, they typically manage, as there are many online resources available for developers on I18N-friendly coding practices. Validation teams are then faced with a major challenge: How to validate that these requirements are properly met and guarantee the global readiness of the product with a high degree of confidence? Testing the proper internationalization of a SW application can easily push your validation team outside of their comfort zone – After all, they might not know what to look for, or where to start.

Using real-life examples, we'll give you a practical and structured overview of the kind of issues localization teams encounter on a regular basis. First, we'll introduce development and validation teams to the main aspects of internationalization (I18N) testing, focusing on language-independent activities. Then, we'll show you how to give an international twist to your peer reviews and code scans, how to run your test plans using pseudo-localized builds (and what to look for), and how to work with international test data (and why). Finally, we'll look into a few issues that require a bit of multicultural awareness, and a basic understanding of foreign languages.

Track 2 - Globalization for Flash/Flex Enterprise Applications

Presenter:

Michael Bridgers
Sr. Software Engineer - Globalization, SAS Institute

Flash/Flex is an important RIA environment, but its support for globalization is still evolving. This presentation will describe SAS's approach for adding globalization support. Topics to be covered will include: Handling the locale and locale chain, Packaging and loading localized bundles, Localized themes, Linguistic collation, Static analysis to check for I18N problems, Deployment issues for modular applications and applications with plug-ins, plus other topics.

Track 3 - Proper Name Transcription/Transliteration with ICUTransforms

Presenters:

Sascha Brawer
Google Inc

Martin Jansche
Google Inc

Hiroshi Takenaka
Google Inc

We describe our experience with a deep localization of Google Maps™, where millions of geographic names from diverse origins had to be represented in several target languages, including Russian, Mandarin, and Japanese. For example, a map of Western Europe on maps.google.co.jp shows Japanese labels for almost all labeled features. We tackle the problem of transliterating from several source languages into several target languages by pivoting through an explicit intermediate phonetic representation. Each transliteration scheme is implemented as a sequence of ICU transforms, reusing a few existing transforms from ICU and CLDR, but consisting mostly of transforms that we wrote

Yui Terashima
Google Inc

specifically for this problem. Dividing the problem this way results in many reusable components that make it simple to transliterate between multiple languages. We discuss the steps that go into building transliteration rules, describe existing official and de facto standards and guidelines, and give suggestions for what to do when no consistent guidelines are available. We provide general recommendations for developing and testing custom ICU transforms.

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

16:00-16:50 **SESSION 5**

Track 1 - ICU Update

Presenters:

Yoshito Umaoka
Software Engineer, IBM

Markus Scherer
Unicode Software Engineer
Google Inc.

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 iPhone or Android 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 4.4, including the latest support for Unicode 5.2 and CLDR 1.8, some improvements for better modularization for mobile phones and other small-footprint devices, faster and flexible Unicode normalization code designed for supporting the Unicode IDNA compatibility processing and other changes (see http://icu-project.org/download/4.4.html). The presentation will also touch on ICU's planned direction for 4.6 and future releases.

Presenter:

Arthur Jin
Senior Program Manager, Microsoft

Track 2 - Microsoft Windows Live Internationalization Framework

Microsoft Windows Live is a suite of web services and client applications designed to keep people connected and make life easier. It currently sim-ships in 48 languages and 76 markets, reaching out to 500 million users worldwide. How is Windows Live internationalized? In this presentation titled "Microsoft Windows Live Internationalization Framework", Arthur Jin, senior International Program Manager at Microsoft talks about how the Windows Live Internationalization Framework drives standards and best practices for product development across Microsoft Windows Live. The presentation will drill into the four basic goals the Internationalization Framework is designed to achieve: consistency, efficiency, scalability and agility which characterizes Windows Live internationalization. Examples, case studies and demos will be presented as appropriate. The goal of the presentation is to share internationalization best practices and challenges and promote advancement of internationalization to benefit our users.

Arthur Jin has 21 years of experience designing software for global use, including fourteen years at Microsoft as an international program manager. In the past four years, Arthur has authored and architected most of the Internationalization Framework which is reflected in the Windows Live product design. Arthur is widely regarded as an internationalization "go to" person in Windows Live product teams.

Presenter:

Richard Sites
Google, Inc.

Track 3 - Statistical Language Detection in Web Pages

Search engine companies prefer to show users pages that they can read. To do this, it is useful to identify the language(s) used on each web page. With billions of web pages in hundreds of languages, an automated statistical approach is needed. In contrast to previous work using short words or groups of three letters (trigrams) to identify perhaps a dozen different languages in single-language well-written text corpi, we look at the more general problem of detecting ~180 languages in ~57 Unicode scripts, in the sometimes mixed-language wild-west text of Web pages. We discuss statistical detection using quadgrams, building statistics offline from the Web itself as corpus, and dealing with unusual pages -- e.g., what goes wrong.

17:00-17:50 **SESSION 6**

Track 1 - Extending ICU with new features

Presenters:

Savithri Jasti

International Components for Unicode (ICU) is a very powerful internationalization software solution, with a flexible and extensible design.

Senior Engineer, Yahoo Inc.

Kirti Velankar
Senior Engineer, Yahoo Inc.

In this session we'll cover the infrastructure of ICU for extending and customizing data and functionality as well as the process of going from requirements to integrating an ICU feature. The addition of select formats to MessageFormat and of new date formats will be used as examples.

Presenter:

Track 2 - How To Achieve World(-Ready) Domination In Silverlight 4

Guy Smith-Ferrier
Internationalization Consultant, Capella Software Ltd.

So you've written your Silverlight application and you want it to work in another language? Then this session is for you. World-Readiness is all of the work that a developer needs to do to globalize an application and make it localizable (i.e. capable of being localized). Whereas these concepts are well established in Windows Forms and ASP.NET, Silverlight is not only a cut-down version of the .NET Framework but also cross platform and client-side. In this session you will learn how to localize Silverlight applications using .resx files, download culture-specific resources on demand so that users only download resources for the culture they need, understand what System.Globalization types and properties Silverlight does not support and why, what globalization and font support you can expect on Windows and the Mac, what the Silverlight installation user experience is for non-English users and what language support you can expect from the Silverlight framework.

Presenter:

Track 3 - Dialect Detection

Khaled Hafez
Google, Inc.

Providing locally relevant content does not only involve providing content that matches the user's language. Certain terms and expressions are exclusively used in some cultures, and those terms makes each culture has its unique dialect. In this presentation we explain an approach for detecting the dialect of an arbitrary piece of text and show how we are going to use this to achieve a new level of local relevance, and thus enhancing the browsing experience of international users on the web.

18:00-20:00 - IUC34 CONFERENCE RECEPTION (IN EXHIBIT AREA)

Wednesday, October 20, 2010

09:00-09:50 SESSION 7

Presenter:

Track 1 - Alert: New Internationalized Domain Names (IDNs) are coming!

Mark Davis
Sr. Internationalization Architect, Google Inc.

Major changes are underway for URLs. Since 2003, internationalized domain names (IDNs) have supported non-ASCII characters in domain names, such as http://ÖBB.at for the Austrian train system. In early 2010, three important events occurred:

- ICANN started allowing top level IDNs, such as in http://президент.рф, so that entire URLs (except for the http://) can be internationalized.
- The IETF issued a revised specification for IDNs
- The Unicode Consortium issued UTS #46, Unicode IDNA Compatibility Processing, with information on handling the new and old versions compatibly.

This presentation puts all of these pieces into context, allowing people to clearly see all of the interrelationships, and how and why the changes are happening. It then focuses on the compatibility and security problems associated with supporting IDNs, presenting concrete approaches (and tools) that software engineers can use to deal with them.

Presenter:

Track 2 - Internationalizing Twitter

Matt Sanford
Tech Manager of International Team, Twitter, Inc

Twitter is growing very quickly and that's even more true outside of the US. In response we have internationalized the service, localized the interface, and even customized Twitter response to existing user behavior that differs from US users.

You'll hear examples tonight from Chile to Japan, and from Unicode support to translation. Like all Twitter features, user input was key from the beginning so we engaged our most passionate bilingual users to help with translating. This required a

community translation tool that would encourage engagement and provide the context that non-professional translators need. Even on a site with as much user generated content as Twitter.com we still had a daunting task to undertake. Some things went well, and some didn't. All of it was a lesson worth sharing.

Presenters:

Martin Benjamin
Kamusi International Project

Deborah Anderson
Researcher, Dept. of Linguistics, UC Berkeley

Laura Welcher Ph.D.

Alan Christy,
Dept. of History, UC Santa Cruz

Track 3 -

Script Encoding Initiative, UC Berkeley by Deborah Anderson, Researcher, Dept. of Linguistics, UC Berkeley

The Script Encoding Initiative began in 2002 as a project in the Department of Linguistics at UC Berkeley. Since its inception, it has assisted in getting over 55 scripts into Unicode. The remaining modern scripts are located in Asia and Africa, while historic scripts are dispersed more widely. About 50 scripts remain to be encoded. Many of the scripts that remain are poorly documented, and involve significant research by script proposal authors. Anshuman Pandey, a graduate student at the University of Michigan who works for SEI, for example, had done considerable work in uncovering information on a number of lesser known scripts in India and other countries in the region, such as Tolong Siki, Tani Lipi and Pau Cin Hau. This presentation will discuss some of the issues (and problems) involved in gathering information for such lesser-known script and will discuss ways to disseminate the information more widely (such as on Script Source, an SIL project and CLDR).

A Translingual Approach to Crowd-Sourcing History: Constructing a Web Archive for Memories of WWII in the Pacific - University of California, Santa Cruz

This presentation discusses an effort to develop a multilingual, transnational, crowd-sourced archive on memories of WWII in the Pacific. A "living" archive for war memories, Eternal Flames will serve as a multilingual research tool and forum for cross-cultural negotiation among war survivors, academics, private scholars and the general public. We hope that our website will help scholars around the world confront major issues facing the global humanities including not only problems of language and cultural barriers, translation, and successful communication between diverse groups in different nations, but also broader issues such as internationalizing research and curricular development, utilizing user created digital sites to conduct meta-research into the nature of the research process itself, and direct investigation into the history-memory conundrum so perplexing to humanities discussions today.

Unicode and CLDR outside of Industry: News from Academic and Non-Profit Projects

The development and widespread use of Unicode and CLDR in industry is well-established, but less well-known are any issues surrounding Unicode and CLDR faced by groups outside of industry. Since the academic and non-profit groups often work “on the ground” in terms of language, script, and locale collection as well as developing language-related digital resources, their perspectives are useful. What types of issues do non-industry groups face in collecting and the posting such data? Are there problems in implementing Unicode and CLDR in language learning and language documentation projects? Is there room for Unicode and CLDR to assist these academic and non-profit groups?

This session will highlight work being done at universities and non-profits. Members of the session will include faculty from UC Santa Cruz, as well as members of the Script Encoding Initiative in the Department of Linguistics at UC Berkeley, the Rosetta Project, and the 100 African Locales Initiative/Kamusi Project International.

The Rosetta Project - Wiki of All Human Language

The Rosetta Project maintains a large digital collection of information about the world’s languages, and as part of the 10,000 Year Long Now Foundation Library, develops projects that experiment with strategies and practices for storing, migrating, and serving this information over long time frames.

This talk will demonstrate a new distributed system that we have developed for serving Rosetta's information about the world’s languages – an archived collection, an open public database, and a wiki that is structured from both of these. The collection of

language information is housed in the Internet Archive, and the database of language information is maintained in Freebase – both the Internet Archive and Freebase are third party resources that are freely available to anyone. The demonstration shows how distributed, free online resources can be harnessed to build, structure and serve large open linguistic datasets.

The Language Commons, a new international consortium that supports open datasets for all the world's languages, has recently adopted the Rosetta distributed archive and wiki as its model for storing and serving language resources and datasets. With one page for every human language, the wiki would be an ideal place to track the linguistic information needed to support Unicode proposals.

100 African Locales Initiative/Kamusi Project International

The 100 African Locales Initiative was an effort by the African Network for Localization (ANLoc) to ignite IT development for Africa, where one billion people speak 2000 languages, but few Unicode-CLDR locales existed by 2008. A locale is essential for IT localization for a language. Without the services locales enable, Africans are limited to accessing technology through the former colonial languages. ANLoc sought to democratize access to IT by opening the door for the majority of Africans to use technology in their own languages. The initiative was built on: (1) A user-friendly tool, built by IT46, that simplifies compiling and submitting data; user experience revealed improvements to create locales for many more languages around the world, should interest suffice. (2) Extensive use of networking to recruit and retain participants from linguistic communities, managed by Kamusi Project International; social networks were not magic, however, especially for languages in areas with complicated politics. From 250 languages with over 500,000 speakers, significant data were collected for 72, of which 54 were incorporated in Unicode-CLDR 1.8. Finding volunteers and working with them through completion yielded experiences that will inform additional African localization activities, and produced substantial data now incorporated within CLDR for use throughout Africa.

10:00-10:50 SESSION 8

Presenters:

Martin J. Dürst
Professor, Aoyama Gakuin University

Addison Phillips
Globalization Architect, Lab126

Track 1 - IRIs Beyond the Napkin: A Survey of Internationalized Resource Identifier Issues and Implementation

If the Latin Alphabet is not your (or your customer's) main script, there are many good reasons for including non-Latin characters in a Web address (URL/URI). This presentation will tell you why, when, and how you can and should do this, and provide the necessary background to make things work for servers and clients.

Non-ASCII characters have been used in Web addresses for more than a decade. Such Web addresses have been called Internationalized Resource Identifiers (IRIs), and since 2005 have been specified in RFC 3987. Early this year, the IETF chartered a Working Group to update the RFC 3987.

The presentation will first explain the basic rules for working with IRIs, in particular the conversion to URIs via UTF-8 and percent-encoding. To provide a deeper understanding, we will then concentrate on the major issues that the IRI Working Group is working on addressing:

- Moving from defining IRIs as a presentation element, while restricting protocols to using URIs, to defining IRIs as protocol elements on par with URIs.
- Balancing between syntactical uniformity for long-term simplicity and backwards conformance with established browser behavior in particular for the domain name and fragment identifier parts of an IRI.
- Moving the specification from a before-after descriptive style to a more procedural style that covers edge cases of implementations existing in the wild.
- Comparing, normalization, and security issues for IRIs.
- Restrictions and display advice for bidirectional IRIs.

Presenters:

Jungshik Shin
Software Engineer, Google, Inc.

Track 2 - Chrome Internationalization

A web browser is arguably the most demanding client application in terms of I18N needs. Google Chrome is a multi-platform WebKit based browser and is localized to more than 50 languages. To meet I18N and L10N needs we used ICU for basic Unicode support,

Nebojša Ciric
Software Engineer, Google, Inc.

character encondings, IDN, formatting etc. We also devised a new method for multi-platform resource handling. In addition, Chrome extension infrastructure comes with simple I18N/L10N support such as message replacement and direction detection.

This presentation covers the usage of ICU, Compact Language Detector, and Hunspell in Chrome, the overview of Chrome extension I18N and resource handling, and the way our work on Chrome improved ICU and CLDR data.

Presenters:

Martin Benjamin
Kamusi International Project

Deborah Anderson
Researcher, Dept. of Linguistics, UC Berkeley

Laura Welcher Ph.D.

Track 3 - Unicode and CLDR outside of Industry: News from Academic and Non-Profit Projects (Cont'd)

10:50-11:10 - Morning Refreshments

11:10-12:00 **SESSION 9**

Presenter:

Adil Allawi
Technical Director, Diwan Software

Track 1 - Unicode Fonts for the Desktop, the Mobile and the Web

When Apple asked me to improve the default Arabic font for their latest systems, the requirements were daunting to say the least. The design needed to be relevant for user interface, printing and small screens. It had to support the full Unicode Arabic range, which meant adding an extra 1500 glyphs together with the relevant tables for Arabic shaping, ligatures, justification and kerning.

This presentation covers my approach based on using Unicode data to automate the font creation. I will also examine the challenges faced when applying this data which is more geared to text processing than font development.

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 fonts as they make their way into open standards and become part of the content of the worldwide web.

Presenter:

Michael Kaplan
Program Manager Company, Microsoft

Track 2 - Windows 7 Language Support — How Does it All Fit Together

Microsoft's Windows 7 has 36 localized builds and 50 plus language interface packs (LIP),and supports 100's of different languages with new keyboards, fonts, and Unicode properties. The localized builds can come in many flavors -- Starter Edition, Home Basic, Home Premium, Business, Enterprise, and Ultimate. Besides the localized versions of Windows 7, there is also the support for creating and displaying content in many different languages. This presentation will sort out the different types of and levels of language support that can be found in each of these versions and how they all relate to each other. As a bonus, there will be a quick peek at how the (already shipped) .Net Framework 4.0 (finally!) has parity with language support with Windows 7, and gives a kind of roadmap for how things will likely be kept in sync in the future.

Presenters:

Mohamed Eldawy

Track 3 - Tashkeel -- A library for automatically adding diacritics to undiacritized Arabic text

Software Engineer, Google, Inc.

Mohamed Taha
Software Engineer, Google, Inc.

Arabic diacritics represent short vowels, and are usually omitted in writing since they are inferred from context. This poses a challenge for performing certain text processing tasks, like text-to-speech synthesis and voice recognition. In this talk, we discuss the problem of restoring diacritical marks to languages based on the Arabic alphabet (Arabic, Persian, Urdu... etc). A brief introduction will be given about the Arabic alphabet and the role of diacritics in the way it is written and pronounced. After that, we will present a library that restores the diacritical marks in text missing diacritics. The library is based on a statistical machine learning approach, and is trained from a noisy dataset collected from the web. We will showcase Tashkeel, a product we launched on Google Labs that uses the library to diacritize text and websites. We will also talk about the accompanying API launched on Google Code to enable developers to integrate diacritization into their applications, and discuss possible applications of the library.

12:00-13:00 - LUNCH

13:00-13:50 SESSION 10

Presenters:

Steven Loomis
Software Engineer, IBM

Mark Davis
Sr. Internationalization Architect, Google Inc.

Track 1 - Deploying the CLDR Common Locale Data Repository

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, the latest project status, and how the process is being improved to produce higher-quality data. Ample time will be given for comments and questions from the audience.

Moderator:

Deborah Goldsmith
Senior Software Engineer Apple, Inc.

Track 2 - International Features of the iPhone OS

iOS is the operating system common to the iPhone, iPad, and iPod touch. Currently in its fourth major version, it has a wide array of international features, including localization into 34 languages, rich support for internationalization, and a unique virtual keyboard and text input system with 52 different keyboards. This session covers the international capabilities of the latest version of iOS from both a user and a developer perspective. Topics covered include localization, formatting, text display, and text input.

Presenters:

Richard Ishida
Internationalization Lead, W3C

Aharon Lanin
Software Engineer, Google, Inc.

Track 3 - Extending Bidi Support on the Web

Work has been under way at the W3C to define extensions for handling of bidirectional text on the Web. These improvements draw on the experience of experts who use these languages themselves and have found gaps in the current specifications. Topics under discussion include ways of avoiding directional issues when inserting substrings into text from an external source (such as a database), how to guess which side of a form field to begin displaying text and how to retain information about the directionality of text input into forms, how to ensure that browser chrome respects directional information, how to allow for icons and graphics to be automatically flipped in different directional text, etc. The proposals are targeted initially at HTML5, but are also relevant for other markup. This talk will provide an overview of the proposals made, and progress on their adoption in specifications.

14:00-14:50 SESSION 11

Presenter:

John Emmons
Globalization Architect, IBM

Track 1 - CLDR 1.8 - Meeting the challenges of Africa

The latest version of Unicode's Common Locale Repository (CLDR) version 1.8 incorporates data for 41 new African languages and enhanced data for an additional 13 languages. For this release, the Unicode Consortium partnered with ANLoc, the African Network for Localization, a project sponsored by Canada's International Development Research Centre (IDRC), to help extend modern computing on the African continent. ANLoc's vision is to empower Africans to participate in the digital age by enabling their languages in computers. A sub-project of ANLoc, called Afrigen, focuses on creating African locales.

This presentation will highlight the challenges we faced in collecting and integrating language data from the Afrigen project into CLDR, as well as presenting ideas for the future direction of the CLDR / Afrigen partnership.

Presenters:

Markus Scherer
Unicode Software Engineer, Google Inc.

Katsuhiko Momoi
Staff Test Engineer & I18n Consultant, Google Inc.

Yasuo Kida
Apple, Inc.

Mark Davis
Sr. Internationalization Architect, Google Inc.

Track 2 - Emoji in Unicode 6.0: From Proposal to Deployment

"Emoji" symbols or "picture characters" are widely used in email by more than 80 million Japanese cell phone users. Many Emoji symbols are used as parts of text, as nouns, adjectives, etc. ("I go <ski>"/"I'm <happy>"/"<screaming in fear>") The cell phone systems encode them as characters, via vendor-specific extensions of the Japanese character sets, but there was no standard way to represent them in Unicode.

Unicode 6.0 adds more than 600 symbols to allow general interchange without losing or corrupting data. This also makes sophisticated processing possible, such as text search treating a heart symbol as a synonym for the word "heart". This presentation summarizes the repertoire of Emoji symbols, and strategies for deployment and usage.

Presenters:

Murray Sargent III
Partner Software Design Engineer, Microsoft

Ayman Aldahleh
Software Development Engineer, Microsoft

Track 3 - Tailoring the Unicode Bidi Algorithm

The Unicode Bidi Algorithm (UBA) is a very useful, general, and standard approach for displaying text that contains right-to-left scripts. But there are situations in which it is awkward to use and/or is visually confusing. In such cases, tailoring the UBA can improve the result.

Specifically, in math zones, neutrals other than the period and comma should have the directionality of the math zone. The neutrals include math operators and spaces. An alphanumeric span of characters is displayed according to the UBA, but the span as a whole is treated as object with the directionality of the math zone.

When bidi characters appear in an IRI, strange display may occur. A more readable display results if IRI delimiters have the same directionality as the paragraph (or embedding). Alphanumeric spans in the IRI are displayed in the order determined by the UBA.

According to the UBA, there are cases for which both parentheses of a parenthesized expression have the same glyph. We can fix such parenthesized text displays by ensuring that both parentheses of a matched pair have the same bidi level and that the pair's contents has bidi level(s) greater than or equal to the parenthesis level.

14:50 – 15:10 - Afternoon Refreshments

15:10 - 16:00 SESSION 12

Track 1 - Solving Problems with Locale/Language Identification

Presenters:

Yoshito Umaoka
Software Engineer, IBM

Mark Davis
Sr. Internationalization Architect, Google Inc.

It is crucial for software to understand and communicate linguistic preferences, such as a user's language. For software to interoperate, language and linguistic preferences must be clearly and unambiguously identified. The IETF BCP47 language tag is the core standard for identifying a language of information or representing linguistic preferences, required by HTML, XML, and other modern standards. Using BCP47 may look easy, but it's actually not! For example, the language tags "no" (Norwegian), "nb" (Norwegian Bokmål) and "nn" (Norwegian Nynorsk) are all valid and widely used in Web browsers. But much software fails to recognize the relations between them, and fails to retrieve a Norwegian document tagged by "nb", when "no" is requested. As another example, the Java Locale is stuck on an old version of the predecessor of BCP47, presenting many difficulties for developers.

Unicode defines stable identifiers for languages and locales defined by UTS#35 Unicode Locale Data Markup language(LDML). The Unicode language/locale identifier is based on

BCP47 language tag, but provides some necessary extensions, as well as data and guidelines to resolve many implementation issues, such as picking the best language from those supported by an application. LDML specifies richer locale information with the BCP47 extension "u", so that such information can be communicated via standard protocols. For example, a BCP47 language tag "ja-JP-u-ca-japanese" represents a locale "ja-JP" with calendar ("ca") of type Japanese Imperial ("japanese").

In this presentation, we'll walk through various problems with language/locale identifiers and discuss the solutions provided by Unicode CLDR. We'll also look at the Java Locale, discussing work-arounds for the current situation, the design issues for adopting BCP47 language tags, and the solutions proposed by OpenJDK Locale Enhancement project.

Presenter:

Track 2 - Designing Global Web Service APIs

Norbert Lindenberg
Internationalization Architect, Yahoo! Inc

Can I limit my web service to support only UTF-8? Which parameters do I need to support proper internationalization? Should I use the Accept- Language header or a language parameter in the query string? Can I leave sorting up to the clients of my web service? Which text should I localize, and which should I leave alone? How should I negotiate languages when my web service calls another web service?

Designing web service APIs that support global web applications requires answering numerous internationalization questions. This presentation will discuss some of the trade-offs and solutions.

Presenter:

Track 3 - A Beginner's Guide to Indic Scripts

Christopher J. Chapman
Principal Engineer Monotype Imaging Inc.

Starting a project for the Indian market? Overwhelmed by all the different Indic scripts? Confused by how Indic characters have different glyphs in different contexts? Wondering how Indic fonts are structured? Just curious about Indic scripts in general?

In this talk I will introduce you to the major scripts of India, and show you that there is a system common to all Indic scripts, regardless of differences in appearance. That common basis is reflected in the Unicode encoding for each of the scripts as well as in the OpenType font tables for those scripts. I will describe the common structure of the scripts, and then explore some of the differences between the scripts that you should be aware of when developing software and fonts for these scripts.

16:10 - 17:00 SESSION 13

Presenters:

Track 1 - Address and Phone Number Internationalization - Standards, Technologies and Best Practices

Shaopeng Jia
Software Engineer, Google

Clint Yang
Software Engineer, Google

Lara Rennie
Software Engineer, Google

Addresses and phone numbers are crucial pieces of user information that need to be parsed and formatted correctly for users all over the world. In this talk, Google engineers will discuss Google's techniques for dealing with this challenging topic. We will start with a brief introduction to international address and phone number standards, and the problems/characteristics of addresses and phone numbers globally. After that, we will present Google's approach to formatting, parsing and validating international addresses and phone numbers. The talk will conclude with a demonstration of the newly open-sourced International Phone Number Library, showing how Google's open-sourced phone number technologies could be easily used to benefit developers at large.

Presenter:

Track 2 - Design Considerations for Chinese eCommerce Web Sites

Tex Texin
Xen Master, XenCraft

China is the third largest economy in the world. Owing to its rapid growth, it will overtake second place Japan soon. In 2009, China's $4.9 trillion GDP closed in on Japan's $5.27 trillion. With a growth rates of approximately 9%, China is likely to overtake Japan which is growing more slowly at 4.6%.

China is a factor influencing almost all types of business around the world, both as a producer and as a consumer. China must be considered for its potentiality as a supplier and a market opportunity. Modern business software should be designed to support commerce with China.

The software modifications needed to satisfy the requirements imposed by Chinese language and culture can be significant and time consuming. The development cost can be reduced by designing in the capabilities early in the development effort rather than implementing changes when the demand has made the requirements an immediate need.

This paper surveys the capabilities that software simultaneously supporting both the Chinese market and Western markets (Europe and the Americas) should address. The paper is intended for managers and business executives and provides information to help with planning development and assessing market readiness.

Presenter:

Track 3 - The Past, Present, and Future of Tamil in Unicode

Michael Kaplan
Program Manager
Company, Microsoft

The encoding of Tamil within Unicode has been the subject of concern by many interested parties such as the government of Tamil Nadu for as long as it has been there. It has led to a proposal (built up over the last decade) to try to change the way that Unicode looks at Tamil, as a reflection of the way Tamils look at the language themselves. The culmination of this is an official acknowledgement of Unicode as the way to encode Tamil that is being released this very year. The broader issues of the view of languages and the "rights" of language owners will also be discussed in this case study of a language that has been both wronged and righted as few others have in modern times.

Program is subject to change.

- To Register for IUC34: http://www.unicodeconference.org/registration.htm
  Or, contact Maribeth Mahoney 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.