conference program
speakers' bios
advisory committee
review committee
send me more info
press room
IUC 29 sponsors
exhibits

Program

Monday, March 6

09:00-09:45 **MORNING TUTORIALS**

Presenter:
Asmus Freytag
President
ASMUS, Inc.

Track 1: Unicode 5.0 Tutorial: Part 1 - Characters in Action
Part I of the Unicode 5.0 Tutorial is a uniquely accessible and entertaining way of visualizing the core concepts of the Unicode standard. In this part you will find answers to these questions: What is a Unicode character and how are Unicode characters represented and used in a modern computing environment? How are Unicode characters entered into and displayed on a computer? How are Unicode characters interchanged? What is the interaction between Unicode and rich text (markup)? How do end-users experience Unicode? Throughout Part I, the Unicode 5.0 Tutorial gives typical examples of how the Unicode Standard interacts with the other elements of an internationalized software architecture. With the help of concrete scenarios for the use of Unicode characters you will become familiar with the role the Unicode Standard plays and the benefits of supporting it. Part I of the tutorial provides a concrete context to which the more systematic and detailed treatment of the features of the Unicode Standard presented in Part II and Part III can be related.

Presenter:
Addison Phillips
Internationalization Architect
Yahoo!

Track 2: Internationalization: An Introduction
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 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. Some of the topics covered will include: character encodings and Unicode; processing text in different languages; 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:
Tex Texin
Internationalization Architect
Yahoo!

Track 3: 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.

10:30-10:45 **Morning Refreshments**

09:45 – 12:15 **MORNING TUTORIALS**

Presenter:
Richard Ishida

Track 1: An Introduction to Writing Systems & Unicode
The tutorial will provide you with a good understanding of the many unique

Internationalization Activity Lead
W3C

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.

Track 2 - Internationalization: An Introduction (Cont'd)

Track 3 - Web Internationalization - Standards and Best Practices (Cont'd)

12:30-13:15 **LUNCH**

13:30-15:00 **AFTERNOON TUTORIALS**

Session 1

Presenter:
Asmus Freytag
President
ASMUS, Inc.

Track 1 - Unicode 5.0 Tutorial: Part 2 - Fundamental Specifications
Part II of the Unicode 5.0 Tutorial builds on the concepts introduced in Part I and systematically presents the details of fundamental specifications that are part of the Unicode Standard. Topics include: organization of the Unicode code space; principles used to allocate and unify characters; encoding forms including definition of UTF-8, UTF-16, UTF-32 and when to use each; how to use byte order mark; combining characters and equivalent code sequences equivalent; format characters and other special characters and code points; organization of the Unicode Standard. Part II of the Unicode tutorial is recommended for anyone interested in a systematic overview of the key aspects of the standard. Detailed technical or programming experience is not required.

Presenter:
Deborah Goldsmith
Sr. Software EngineerApple Computer, Inc.

Track 2 - The Dao of Unihan
Over half of the characters in the Unicode Standard are ideographs. This ideographic repertoire, termed Unihan, is intended to provide complete coverage for all the characters in current or past use in all varieties of Chinese, Japanese, Korean, and Vietnamese. In this talk, we will give an overview of the structure of the current repertoire of Unihan and its organization. We will discuss some practical implementation issues and how to deal with them. We will also provide an overview of the Unihan database. This is a large body of normative and informative data which is maintained by the Unicode Consortium and included among the data files which are a part of each release of the standard. We will discuss the nature of the data in the database and how it can be used.

Presenters:
Frank Yung-Fong Tang
Software Engineer
Google

Felix Sasaki
Internationalization Activity Member
W3C

Liam Quin
XML Activity Lead
W3C

Track 3 - Internationalization Features in XPath, XQuery and XSLT
In recent years, the W3C has worked on 17 (!) documents which deal with the XML query language "XQuery" and the transformation language for XML documents "XSLT 2.0", henceforth noted as "QT". The newest QT working drafts include several features for Unicode processing and for general software internationalization. This tutorial discusses the use of QT technology in the context of global (Web) application development. Since XQuery is built on top of the features available in XPath, this tutorial will start with an introduction to XPath 2.0. We will introduce the basics of XQuery to the audience by showing how to use the FLWOR expression to query on an XML document. We then focus our discussion on the Regular Expression and Collation facility in XPath/XQuery. We will also briefly describe some current work and issues in the research of developing XQuery Full Text Extension. XSLT 2.0 is - like XQuery 1.0 - based on XPath 2.0. We will describe the common properties of XSLT 2.0 and XQuery 1.0, focusing on XML Schema datatypes, in specific the types for dates and time zones. XSLT-specific features for internationalization will also be covered.

15:00-15:15 **Afternoon Refreshments**

15:15-17:00 **AFTERNOON TUTORIALS**

Session 2

Presenter:
Asmus Freytag
President
ASMUS, Inc.

Track 1 - Unicode 5.0 Tutorial: Part 3 - Unicode Algorithms
The Unicode Standard and related specifications by the Unicode Consortium specify a number of algorithms. The specification of these algorithms in the Unicode Standard depends on the Unicode Character Properties. Part III of the Unicode 5.0 Tutorial surveys the algorithms specified in the Unicode Standard, and extends the discussion of Unicode character properties as they relate to each algorithm. Part III covers many general aspects of Unicode algorithms: Unicode Algorithm and the difference between an abstract algorithm from an actual implementation; relation between algorithms and Unicode Character Properties; techniques to access character properties. Several algorithms are discussed in more detail for example: Unicode Normalization and the requirements it addresses, including a discussion of the Unicode Normalization forms NFC, NFD, NFKC, NFKD, their interaction with the Web and what programmers need to know in applying normalization; the Unicode Bidirectional Algorithm, and its interaction with text layout; text boundary determination and character foldings and much more. Part III of the Unicode 5.0 Tutorial is more detailed and will touch on the description of algorithms and other material that may require some familiarity with technical concepts.

Track 2 - The Dao of Unihan (Cont'd)

Presenters:
Steven R. Loomis
Software Engineer
IBM Corp.

Vladimir Weinstein
Software Engineer
IBM Corp.

Track 3 - Advanced ICU Topics
ICU is a mature, widely used set of C/C++ and Java libraries for Unicode support and software internationalization and globalization (i18n/g11n). It grew out of the JDK 1.1 internationalization APIs (which the ICU team contributed) and continues to be developed for the most advanced Unicode/i18n support. ICU is widely portable and gives applications the same results on all platforms and between C/C++ and Java software. This tutorial walks the audience through the core concepts of using the ICU library, providing an introduction to setup and use of ICU in practice. The concepts are presented via a concrete internationalization task, illustrating the use of ICU for character conversion, collation, message formatting and text boundary analysis. The tutorial will walk through code snippets to solve this task, followed by demonstration applications and discussion of core features and conventions, advanced techniques and how to obtain further information.

17:00 -18:00

Session 3

Presenter:
Asmus Freytag
President
ASMUS, Inc.

Track 1 - Unicode 5.0 Tutorial: Part 3 - Unicode Algorithms (Cont'd)

Presenter:
Thomas Milo
Director, DecoType,
The Netherlands

Track 2 - Abridged Arabic Tutorial

This new tutorial is completely redesigned to provide a global conceptual framework, while at the same time serving as an introduction for the uninitiated. It will cover the structure of Arabic script; historical origin; graphic assimilation; horizontal and vertical connections; the calligraphic dimension; script analysis & synthesis; font technology; graphemes, transcription & transliteration; what to encode; ambiguous letters and codes; code page legacy; compatibility; standard vs. qur'anic orthography; ambiguity in character ordering; and types of line-breaking.

19:30 -21:00 **BIRDS OF A FEATHER**

Presenter:
Addison Phillips
Internationalization Architect
Yahoo!

Track 1 - Language and Locale Identifiers
This BOF session allows participants interested in language identifiers ("language tags") and locale identifiers to discuss the latest issues involved, including: RFC 3066 and its successor; the differences between locale and language identifiers; and best practices for these identifiers.

Presenter:
BOF Leader TBA

Track 2 - Unicode Committee
This BOF session gives participants the opportunity to meet some of the members of the Unicode technical committees, to find our more about how they work and what issues they face. The Unicode Technical Committee is the committee responsible for the Unicode Standard, and related software globalization standards and documents; the CLDR Technical Committee is responsible for locale data, and related localization standards and documents.

Presenter:
Erkki I. Kolehmainen
Coordinator, Cultural Diversity Issues in ICT Research
Institute for the Languages of Finland (RILF)

Track 3 - Design Principles for A Regional, Multilingual Keyboard
How to provide an open-ended, international "Unicode" keyboard layout for a maximum repertoire with a minimum number of pre-allocations, yet tailored for the particular region.

Tuesday, March 7

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

Mark Davis - **President, Unicode Consortium**

09:15-10:00 **KEYNOTE - Tuoc Luong, Executive Vice President, Engineering & Technology, Ask Jeeves, Inc.**

Going Global with a Search Engine

Taking a web service global is best done from the beginning. However, often is the case that software service is designed adn implemented for the United States first then re-designed later for other languages and markets worldwide. This is the case with the Ask Jeeves search service. This talk will touch on all issues (both technical and non-technical) dealing with such a re-design and move globally. Technical issues from language identification, segmentation to geographical latency. Non-technical issues from team dynamics, process in transition to business decisions. The talk will give the audience a good flavor of the issues involved in moving a highly scaleable search engine internationally

10:00-20:00 **EXHIBIT AREA OPEN**

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

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

Presenter:
Doug Barbin
Director-Compliance Solutions
VeriSign, Inc.

Track 1 - The goal of this session is to compare what is happening online today to classic scams involving identify theft, fraud, and corruption. Lessons that have been learned -- and those which clearly have not -- will be discussed. What companies and individuals are doing to protect themselves will be reviewed, as well as what actions should be taken if digital fraud occurs. Case studies will be presented.

Presenter:
Murray Sargent III
Sr. Software Design Engineer
Microsoft Corporation

Track 2 - Editing and Display of Mathematical Text using Unicode
This talk describes and demonstrates how Unicode's rich mathematical character set combined with OpenType font technology, TeX's mathematical typography principles, and enhanced autocorrection can be used to produce high-quality, streamlined technical text processing.

Presenter:
Mark Davis
President,
UNICODE Consortium

Track 3 - Latest News in Globalization
This presentation provides an update on the latest developments in software globalization from the Unicode Consortium, summarizing the most important changes in the Unicode character encoding standard, related globalization standards and specifications, the locale data repository (CLDR), etc. It also

[Top]

describes important related developments from the IETF, ICANN, the W3C, and others.

11:30-12:20 **SESSION 2**

Presenter:
Mark Davis
President,
UNICODE Consortium

[Top]

Track 1 - Unicode Security
Because Unicode contains such a large number of characters and incorporates the varied writing systems of the world, incorrect usage can expose programs or systems to possible security attacks. This presentation describes some of the security considerations that programmers, system analysts, standards developers, and users should take into account, and provides specific recommendations to reduce the risk of problems. It also describes the mechanisms recommended by the Unicode Consortium and others for dealing with these issues.

Presenter:
Marc Durdin
Director
Tavultesoft Pty Ltd

[Top]

Track 2 - Keyman: A New Way of Thinking about Keyboard Input
Despite advances in every area of computing, the keyboard has changed little since the typewriter. Existing solutions for multilingual input have tended to use simple key mapping, with nonintuitive dead keys and complex modifier key combinations. In order to have ideal text input, keyboard input methods need to provide validation, reordering and Unicode normalisation. Keyman is a tool which makes this possible through a contextual input mechanism. Contextual input also allows both logical and physical ordering of input, regardless of the underlying storage order, and phonetic or alternate script input.

Presenter:
Addison Phillips
Internationalization
Architect
Yahoo!

[Top]

Track 3 - Language Tags and Locale Identifiers
The recently approved RFC 3066bis updates the most common standard for language identification. These changes allow for a more regular, structured set of tags. Understanding these changes is the key to realizing their potential and implementing solutions that use them. Language tags and related work on locale identifiers are being used to address internationalization issues ranging from common locale data (CLDR) to Web services internationalization. New standards and specifications are being developed that affect developers and content authors alike. This presentation, from one of the authors of RFC 3066bis, will explore the new tags, their use in various applications, and their relationship to locale identification.

12:30-13:00 **LUNCH**

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

Presenter:
Tex Texin
Internationalization
Architect
Yahoo!

[Top]

Track 1 - Unicode-enabling PHP
Up to now, PHP has provided only marginal multibyte and Unicode support. This session discusses the project to add Unicode support to PHP 5 including incorporation of the ICU library. The changes to the PHP language, potential migration issues and other aspects of the ongoing development will be provided.

Presenter:
Russ Rolfe
Lead Program
Manager
Microsoft Corporation

[Top]

Track 2 - Windows Vista, An Ever Expanding View of Internationalization
Windows Vista expands upon the foundation created in Windows 2000 and Windows to support users' Internationalization needs. In this presentation, we will introduce the newly supported locales, user input options and the extended font coverage. Then we will show how the OS's localized User Interface will be wholly supported with the Multilanguage User Interface (MUI) technology introduced in Windows 2000 and improved in Windows XP. Plus, we will discuss how Language Interface Packs (LIPs) will be used to broaden the localized language coverage. Finally the concept of custom cultures will be presented.

Presenters:
Felix Sasaki
Internationalization
Activity Member

Track 3 - How to Express Information about Internationalization and Localization in XML Documents and Schemata
This presentation describes a first proposal for the Internationalization Tag Set (ITS), which has been developed by the W3C i18n ITS Working Group.

W3C

Richard Ishida
Internationalization Activity Lead
W3C

The purpose of ITS is two-folded: First, ITS provides a set of XML elements and attributes which can be used to express information about internationalization and localization purposes. And second, ITS defines mechanisms on how to express this information within an XML document, a schema document or separately. This presentation focuses on the latter aspect, giving examples from various types of XML data.

Top

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

Presenters:
Naoto Sato
Member of Technical Staff
Sun Microsystems

Craig R. Cummings
Principal Software Engineer
Oracle Corporation

Track 1 - New Internationalization Features of the Java Platform -- Java SE 6
See what internationalization features are planned for the next version of the Java Platform -- Java SE 6 (codename Mustang). Highlighted in this session will be features such as pluggable locale data, new Japanese calendar support, resource bundle enhancements, and normalization.

Top

Presenter:
Marypat Meuli
Lead Program Manager
Microsoft Corporation

Track 2 - Microsoft Office 12 - Internationalization Expanded
In this presentation we will discuss Office 12's broadened support for worldwide users. Newly supported Office locales will be discussed as well as the depth of locale support across Office documents. The language neutral architecture for improved multilingual support and deployment will also be covered. In addition, we will give an overview of some of the new international features in Office 12, including the Language Reference ToolTip and the English Writing Assistant. We continue to expand our work with Language Interface Packs, and will highlight some of the new UI languages which will be supported in Office 12.

Top

Presenter:
John Emmons
Globalization Architect
IBM Corp.

Track 3 - Common Locale Data Repository (CLDR) Overview
Unicode's Common Locale Data Repository is a project whose mission is to provide a general XML format for the exchange of locale information for use in application and system development and a repository of common locale data in that format. This presentation will go into the details regarding exactly what types of locale information are available in CLDR and how this data is intended to be interpreted according to the Locale Data Markup Language specification (Unicode Technical Report # 35). Topics include a discussion of the various locale data types, the locale inheritance model, locale aliasing, CLDR supplemental data and metadata, and the POSIX locale generation tools.

Top

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

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

Presenter:
Deborah Goldsmith
Senior Software Engineer
Apple Computer, Inc.

Track 1 - International Features of Mac OS X
Mac OS X is a modern, robust, Unix-based operating system. This session covers the international capabilities of Mac OS X primarily from an end user perspective, with a particular emphasis on new features in Mac OS X 10.4 Tiger, the latest version. Topics covered include supported languages, input methods and keyboard layouts, locales, font technologies, and user customization. Topics of interest to software developers and language experts will also be considered.

Top

Presenter:
Marin Millar
Globalization Manager
Microsoft Corporation

Track 2 - Building Localized and Globalized Windows Applications with Visual Studio 2005
This talk will cover how to use the features in the .NET Framework 2.0 to build globalized and localizable Windows applications. It will demonstrate new features such as Unicode test processing, custom cultures, localization and Click Once deployment with Visual Studio 2005.

Top

Presenter:
Mark Garrett
Support Developer
ModernGigabyte LLC

Top

Track 3 - A Case Study in Web Internationalization: Using the CLDR Online With PHP
This presentation is a case study of our small Web firm that has transformed our flagship product, ModernBill, from an English-centric product to an internationalized Unicode product. It will include a discussion of the old program architecture and how it hampered internationalization and localization. It will continue with the process of assessing the different options for making the software more locale aware, and the implementation of our final solution. The session will end with the generalized internationalization and localization principles we uncovered and how they apply to other Web developers.

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

Presenters:
David Robinson
Ienup Sung
Nicolas Williams
Sun Microsystems

Top

Track 1 - File Systems, Unicode, and Normalization
When you try to retrofit existing file systems to support multilingual file names in a predictable manner, you encounter various issues that need to be resolved. This technical presentation discusses the issues such as why Unicode has to be the choice of the character set for the file systems, how the traditional non-Unicode codesets should be supported, the role of Unicode normalizations, what would be the performance implication, how to deal with possible failure cases, compatibility and inter-operability with other file systems and protocols. Finally, we will also try to provide a comparison on possible resolutions and outcomes.

Presenter:
Michel Suignard
Senior Program Manager
Microsoft Corporation

Top

Track 2 - IDN Support in Internet Explorer 7
The talk briefly introduces the concept of International Domain Name (IDN), its implementation in the Windows platform library and in Internet Explorer IE7. It also shows how security concerns about spoofing and phishing are addressed, using mitigation strategies inspired from the Unicode Technical Report TR36 (Unicode Security Considerations) and Technical Standard TS 39 (Unicode Security Mechanisms) that the speaker co-authored with Mark Davis.

Presenter:
Weiran Zhang
Development Manager
Oracle Corporation

Top

Track 3 - Client-Tier Globalization - The New Frontier
This presentation discusses client-tier globalization challenges and solutions based on a common deployment architecture using J2EE application with open standard DHTML/JavaScript technology as the front-end. We will examine how to leverage JavaScript and Java technologies to build a complete client-side framework for simplifying the development of rich Web clients to support multiple languages and locales consistently across different browsers. We will illustrate approaches to tackle client-side internationalization issues including locale determination, translation resource management, locale-sensitive data processing, performance, and so on. Demonstrations and code examples are included to showcase concepts and best practices.

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

Wednesday, March 8

09:00-09:45

KEYNOTE - Colonel Daniel L. Scott, Assistant Commandant, Defense Language Institute Foreign Language Center

UNICODE as a "Unifying Force" in Language Education
In the aftermath of 9/11 and the subsequent Global War on Terrorism, a new family of languages has taken center stage. No longer are languages from the Cold War era, Russian, Czech, or German filling the corridors of the Pentagon. No longer are mainstays of the academic world – Spanish, French and Italian – garnering much attention. In this modern world conflict, ironically, it's often the ancient languages that are emerging and important -- languages, which, until quite recently, have not been the recipients of much attention. As a nation, we need to quickly educate and train large numbers of linguists and cultural experts, and we need to use our technology to help us do that. Unfortunately, most of these ancient languages are not ready for prime time in terms of computer support. Many languages are only available in one font face that may or may not render the characters in a legible or correct form. Screens often show a mishmash of partially rendered characters

interspersed with the telltale “squares of death.”

The Defense Language Institute Foreign Language Center mission is to train a new generation of linguists in these languages. We don’t have the luxury of time to do all this work in the classroom with traditional textbooks—we need to reach our students using the new technologies of the web and gaming communities. The DLIFLC has created an abundance of emerging language materials ranging from printed Language Survival Guides to fully interactive, web-based language training delivered on CDs and across the Internet. Regardless of the medium of delivery, the DLI seeks solutions that will allow seamless portability from one operating system or application to another. This is absolutely crucial in meeting the nation’s growing demands for language materials. By establishing and strictly adhering to internal policies regarding the use of UNICODE fonts, DLI can export critical materials throughout the country with total confidence that they function as desired.

More importantly, a successful technical convention such as UNICODE brings about change that is non-technical. UNICODE will help us transform language training by streamlining curriculum, implementing web-based testing, conducting on-line classes, and fielding self-paced study and reference materials for linguists at all proficiency levels and anywhere in the world. Linguists can use these references from their offices, their homes, and from their palm pilot at a checkpoint. In this sense, UNICODE becomes a “unifying force” for language education and support. We support and applaud the efforts of the UNICODE community to lead the way.

09:45-10:00 **Morning Refreshments**

10:00-10:50 **SESSION 7**

Presenter:
Mark Davis
President,
UNICODE
Consortium

Track 1 - Globalization Gotchas: What to Watch Out For
This presentation covers the main pitfalls that all programmers should be aware of, so that they can avoid stumbling into them when developing globalized (internationalized) software products. It covers core areas of Unicode, but also such areas as character conversion, text comparison (sorting/matching), locales, date/time formatting, and many others.

Top

Moderator:
Debbie Anderson
Researcher
UC Berkeley

Top

Track 2 - Panel: Unicode and Academia (continued)
This panel is composed of speakers representing the various Unicode-related projects being done in academia today. The goal of the session is to inform audience members of ongoing work (and challenges faced), and to encourage partnerships between the academic world and Unicode encoders, implementers, and developers so the task of "finishing the job" of encoding the scripts of the world and making Unicode "work" are realized.

Panelists:
Prof. Johannes Bergerhausen, University of Applied Sciences Mainz
Charles Riley, Sterling Memorial Library, Yale University
Richard Cook, Dept. of Linguistics, UC Berkeley
John Hudson, Society of Biblical Literature / Tiro Typeworks

Presenter:
Andrew Heninger
Senior Software
Engineer
IBM Corp.

Top

Track 3 - Unicode Processing with Regular Expressions
This presentation will review the issues and techniques involved in writing Regular Expressions for Unicode data. The guidelines from Unicode Technical Report #18 will be reviewed, including a discussion of Unicode encoding forms, character properties and classes, text boundaries, case sensitivity and normalization, and the implications of all of these for handling different languages in regular expressions. The presentation will also survey the capabilities and limitations of those regular expression implementations known to provide significant support for Unicode.

11:00-11:50 **SESSION 8**

Presenters:
Scott Atwood
Staff Software
Engineer
June Wang
Engineering Manager
PayPal

Track 1 - Migrating PayPal to Unicode: A Case Study
PayPal, with 80 million registered users, is a high volume website that processes millions of financial transactions every day. In 2003 PayPal was only localized for the United States and the United Kingdom and used US-ASCII from the browser all the way to the DB. In 2004 PayPal embarked on an ambitious project to completely revise all string handling to use Unicode end-to-end without disrupting PayPal's business or its terabytes of existing data. This case study will discuss that project from inception to completion,

exploring the challenges we faced, the choices we made both good and bad, as well as the final outcomes. A testament to success of the project is that nobody noticed we did it.

Top

Moderator:
Debbie Anderson
Researcher
UC Berkeley

Track 2 - Panel: Unicode and Academia (Continued)
Panelists:
Prof. Johannes Bergerhausen, University of Applied Sciences Mainz
Charles Riley, Sterling Memorial Library, Yale University
Richard Cook, Dept. of Linguistics, UC Berkeley
John Hudson, Society of Biblical Literature / Tiro Typeworks

Presenter:
Michael Kaplan
Technical Lead
Microsoft Corporation

Top

Track 3 - Sorting It All Out: An Introduction to Collation
In a properly globalized product, users will have properly collated data-e.g., in the file system, in a database, in an e-mail address book. How should implementers go about ensuring culturally-correct collation in a product? What are the basic linguistic issues of collation, and how do they manifest themselves in technology? This presentation will explain the basic tenets of collation in language, debunk some myths about collation in globalized software, show how collation functions are used (using examples from the Win32 API), and touch upon best practices.

12:00-13:00 **LUNCH**

13:00-13:45 **KEYNOTE - Charles Bigelow, Vice President, Bigelow & Holmes Inc.**

The Effect of Unicode on Type Design
The widespread adoption of Unicode has affected type design in ways that were neither anticipated nor intended, but which may become even more significant in the future. The main factor is the creation of large fonts ("large" in the sense of many characters) which incorporate characters for several orthographies, scripts, and symbols. A second factor is the structure of the Unicode Standard, organized by named blocks of orthographies, scripts, and symbols. A third factor is conflict between Unicode's definitional distinction of glyphs from characters, and the naive user's "common sense" view that the glyphs depicted in the Unicode manual are in fact the characters. Finally, there are legibility factors, user-interface issues, and security problems that arise when different characters are represented by glyphs, or combinations of glyphs, that appear similar or identical in some circumstances, especially on display screens at small sizes and low resolutions. This talk will be illustrated by examples of glyphs and fonts from a variety of typefaces, ranging from Herman Zapf's Euler fonts for mathematics, to Arial Unicode and Lucida Grande and other "global" fonts, as they are sometimes termed, as well as scripts from Latin to Arabic to Kanji. The discussion will occasionally use terms and notions borrowed from linguistics - such such as analogy, phonemics, graphemics, grapholects, etc. - to explain and analyze the various factors.

14:00-14:50 **SESSION 9**

Presenter:
Peter Linsley
Sr. International Product Manager
Ask Jeeves

Top

Track 1 - Challenges of Searching the Global Internet
Global Search Engines work hard to deliver on the simple promise to take your keywords, sift through several billion documents and come up with a small handful most likely to answer your query. And all this within a few milliseconds. With a focus on language, character set, and cultural issues, the presentation will cover some key challenges in crawling, indexing and making documents of the world searchable.

Presenter:
Markus Scherer
ICU Team Manager/ Software Engineer
IBM Corp.

Top

Track 2 - ICU Overview: The Open-Source Unicode Library
ICU is the premier Unicode-enablement software library, providing a full range of services for supporting internationalization - especially in server environments. ICU is principally developed by IBM, and used in IBM products, but is also freely available as open-source. It provides cross-platform C, C++ and Java APIs, with a thread-safe programming model. The ICU project is licensed under the X License, which is compatible with GPL but non-viral; it can be freely incorporated into any product. This presentation will provide an overview of ICU, with special emphasis on new features scheduled for the

ICU 3.6 release (2006 Q2), including new tools that significantly simplify modularization and installation.

Presenter:
Michael Kaplan
Technical Lead
Microsoft Corporation

Top

Track 3 - Tales of Incorrect String Comparisons
Over the last few years, awareness of the internationalization support in Windows and the .NET Framework has dramatically increased. Unfortunately, so has unintentional misuse of the available collation and casing features that each platform provides. This talk will give the best practices as told by showing the consequences of doing it wrong. All names will be changed to protect the guilty (or the embarrassed!), but you'll leave this presentation knowing the best way to make use of the collation and casing support in both Windows and the .NET Framework.

14:50-15:10 **Afternoon Refreshments**

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

Presenters:
Pavol Zavarsky
Wunna Ko Ko
Yoshiki Mikami
Yew Choong Chew
Nagaoka University of Technology

Tatsuo Kobayashi
Scholex Co., Ltd

Top

Track 1 - Unicode Spreading on the Web: A Case of Asian Domains
In the paper we present some results from our ''Language Observatory" project. Language Observatory surveys digital language activities on the Internet. The scalable fully-distributed Web crawler developed by our research collaborators allows us to collect, parse and extract meta information from more than 15 million Web pages every day. This paper demonstrates clearly that the Web pages development community in South and South East Asia has been moving fast towards the use of the Unicode standard. The paper shows the trends in spreading of Unicode encoding of languages to the Indian subcontinent and other parts of Asia.

Presenter:
John Emmons
Globalization Architect
IBM Corp.

Top

Track 2 - Migrating C Language Applications to Unicode
In the real world, there are many applications that were written "before the enlightened era of Unicode" and need quite a bit of enhancement to bring them up to present day best practices regarding use of Unicode and globalization in general. This presentation will discuss the challenges that system designers and architects face when migrating a legacy C language application to use the Unicode Standard, and will offer practical suggestions on the types of tools available to programmers to make this transition.

Presenters:
Eric Mader
Andrew Heninger
Senior Software Engineer
IBM Corp.

Top

Track 3 - Character Encoding Detection
Charset detection allows text data in an unknown encoding to be read and understood. Charset detection is commonly used in Web browsers to correctly display pages even when the author neglected to include a charset declaration. This presentation will discuss the nature of charset detection and the situations where it can be used. An overview of current techniques will be presented. The latest release of the ICU Unicode support library includes a new charset detection service. The approaches used and results obtained by ICU will be described.

16:10-17:00 **SESSION 11**

Presenters:
John O'Neil
Language Technology Architect
Thomas Emerson
Senior Software Architect
Basis Technology

Top

Track 1 - Large Corpus Construction for Chinese Lexicon Development
The World Wide Web provides an important source of natural language data in many languages. However, it doesn't include annotation about linguistic structure, so it's necessary to use very large corpora to infer it. We developed a system for continuous, automatic acquisition of a Chinese lexicon. An up-to-date lexicon is needed for many applications, but Chinese is written without spaces between words, so determining word boundaries is the primary problem. We discuss our experience with using the Chinese Web for lexicon construction, focusing on both low-level details and problems we experienced during our initial proof-of-concept experiments, and on algorithmic issues.

Presenter:
Doug Felt
Technical Lead - ICU for Java
IBM Corp.

Top

Track 2 - ICU in Eclipse
ICU is being made available in Eclipse through the introduction of the com.ibm.icu plugin that provides access to the ICU4J jar. ICU's globalization enhancements and support for CLDR resource data greatly enhances the locale support available to developers. The Eclipse development platform itself will make use of the plugin to enhance its UI. This paper will discuss the features that ICU provides to all developers, and will also describe how Eclipse was modified to use them.

Presenter:
Raghuram Viswanadha
ICU Team
IBM Corp.

Top

Track 3 - StringPrep: Unicode in Network Protocols

Network protocols require consistent comparison of strings. The StringPrep framework (RFC 3454) facilitates this function. It provides sets of rules that can be applied to strings to prepare them for use in any protocol or program. Each system sets up a profile of StringPrep by selecting a set of rules. Important profiles such as NamePrep, NFS, ResourcePrep, NodePrep are explained. The usage of StringPrep and IDNA frameworks is illustrated by implementation in International Components for Unicode (ICU).

Program is subject to change.