What's new in ICU 4.2

Steven R. Loomis
IBM

Markus W. Scherer
Google

33nd Internationalization and Unicode Conference

San José, California, USA • 2009 October 15th

Tuesday, November 3, 2009

ICU (International Components for Unicode) is an open source development project sponsored, supported, and used by many organizations. It is dedicated to providing robust, full-featured, commercial quality, freely available Unicode-based technologies. Comprehensive support for the Unicode Standard is the basis for multilingual, single-binary software. ICU uses the most current versions of the standard, and provides full support for supplementary characters.

As computing environments become more heterogeneous, software portability becomes more important. ICU lets you produce the same results across all the various platforms you support. It offers great flexibility to extend and customize the supplied system services.

For more information, see the ICU website: http://icu-project.org

Agenda

- Isn’t Unicode enough?
- Why ICU?
- Where is ICU?
- What’s new in ICU?
- What’s next for ICU?

Agenda

- Isn't Unicode enough?
- Why ICU?
- Where is ICU?
- What's new in ICU?
- What's next for ICU?

Isn’t Unicode Enough?

- Handles all modern world languages
- Efficient and effective processing
- Lossless data exchange
- Enables single-binary global software
- But… all languages ⇒ large, complex standard
  - 1,400 pages + Annexes + additional standards
  - More than 100,000 characters
  - Major update every 3 years; minor update about once a year
  - 80+ character properties, many multi-valued
  - Affects many processes: display, line-break, regular expressions, …

Tuesday, November 3, 2009

Unicode (and the parallel ISO 10646 standard) defines the character set necessary for efficiently processing text in any language and for maintaining text data integrity. In addition to global character coverage, the Unicode standard is unique among character set standards because it also defines data and algorithms for efficient and consistent text processing. This simplifies high-level processing and ensures that all conformant software produces the same results. The widespread adoption of Unicode over the last decade made text data truly portable and formed a cornerstone of the Internet.

Unicode enables lossless exchange of multilingual data between different types of computing systems, as well as single-binary installations of software which can handle text in all languages.

As a result, the Unicode Standard is complex and voluminous and not trivial to implement. ICU supports all Unicode characters, is regularly updated to the latest Unicode version, and implements and provides most of the properties and algorithms.

Internationalization, Localization & Locales

- Requirements vary widely across languages & countries
  - Sorting
  - Text searching
  - Line breaks
  - Bidirectional text processing and complex text layout
  - Date/time/number/currency formatting
  - Codepage conversion
  - …and so on
- Performance is key
  - It is easy to do the right thing
  - It is hard to do it fast

Tuesday, November 3, 2009

The design and architecture of software that can work with multiple languages and cultural specializations is called internationalization. It involves taking into account a variety of attributes for many areas of text handling and data input and output. The most common attributes are the written language and the country or region for which data is processed or presented. Standard codes for these attributes, and sometimes others, are often combined into "locale identifiers". Depending on the context, the term "locale" refers either to such locale identifiers or to the relevant collection of associated data and behaviors.

In addition to these familiar attributes, others are also important and cannot be reliably inferred. For example, currency codes and codepages need to be identified reliably for correct results.

Localization provides internationalized software with locale-specific User Interface elements (text, images, layout) and sometimes functionality for regional business rules or similar.

The term globalization is sometimes used as a synonym for internationalization. We use it more narrowly, for software which can be compiled and installed once and handles text in all languages at the same time, as opposed to requiring recompilation for each locale. Such software needs to use Unicode for text processing.

It is often relatively easy to satisfy the requirements from one language or culture, or a small number of closely related ones. However, with the diversity of requirements from many languages and cultures on many processes, and the desire for high performance in many cases, the implementation of these processes can become rather complex. The ICU libraries provide "shrink-wrapped", reusable, tested implementations that were designed with performance in mind.

Agenda

- Isn't Unicode enough?
- Why ICU?
- Where is ICU?
- What's new in ICU?
- Q and A

ICU Features

- Unicode text handling
- Charset conversions (175+)
- Charset detection
- Collation & Searching
- Locales from CLDR (450+)
- Resource Bundles
- Calendar & Time zones
- Complex-text layout engine
- Unicode Regular Expressions
- Breaks: word, line, …
- Formatting
  - Date & time
  - Durations
  - Messages
  - Numbers & currencies
  - Plurals
- Transforms
  - Normalization
  - Casing
  - Transliterations

Tuesday, November 3, 2009

In addition to basic Unicode standard conformance, both the ICU Java library ("ICU4J") and the C/C++ libraries ("ICU4C") also provide a full set of internationalization features listed above.

Notes on C/C++ vs. Java

•ICU C/C++ and Java APIs do differ slightly due to the differences of programming languages. Sometimes the feature development in ICU4C leapfrogs ICU4J or vice versa by 1-2 releases.

•Since ICU is open source and closely tracks the Unicode Standard, ICU can support changes and additions to the Unicode Standard much more quickly than Java. Java support for Unicode is tied to major releases of the JDK, and can lag the Unicode Standard by a year or more.

ICU Works Everywhere

- International Components for Unicode
- Globalization / Unicode / Locales
- Mature, widely used set of C/C++ and Java libraries
  - Basis for Java 1.1 internationalization, but goes far beyond Java 1.1
- Very portable – identical results on all platforms / programming languages
  - C/C++ (ICU4C): 30+ platforms/compilers
  - Java (ICU4J): IBM & Sun JDK
- Full threading model
- Customizable & Modular
- Open source (since 1999) – but non-restrictive
  - Governed by a Project Management Committee
  - Contributions from many parties

Tuesday, November 3, 2009

International Components for Unicode (ICU) is a mature set of widely used C/C++ and Java libraries. They are portable to many environments and platforms. There are 3 sub-projects of ICU. There is ICU4C, which is written in C and C++. There is ICU4J which is written in Java.

Mature: celebrated 10 years this year.

ICU is distributed under the X license. The license allows ICU to be incorporated into a wide variety of software projects using the GPL license, while also allowing ICU to be incorporated into non-open source products. You can read the license on ICU's web site for details.

Agenda

- Isn't Unicode enough?
- Why ICU?
- Where is ICU?
- What's new in ICU?
- What's next for ICU?

Where is ICU?

ABAS Software, Adobe, Amazon (Kindle), Amdocs, Apache Xalan XSLT, Apache Xerces XML, Appian, Apple, Argonne National Laboratory, Avaya, BAE Systems Geospatial eXploitation Products, BEA, BluePhoenix Solutions, BMC Software, Boost, BroadJump, Business Objects, caris, CERN, Debian Linux, Dell, Eclipse, eBay, EMC Corporation, ESRI, Free BSD, Gentoo Linux, Google, GroundWork Open Source, GTK+, Harman/Becker Automotive Systems GmbH, HP, Hyperion, IBM, Inktomi, Innodata Isogen, Informatica, Intel, Interlogics, IONA, IXOS, Jikes, Mathworks, Mozilla, Netezza, OpenOffice, Lawson Software, Leica Geosystems GIS & Mapping LLC, Mandrake Linux, OCLC, Progress Software, Python, QNX, Rogue Wave, SAP, SIL, SPSS, Software AG, Sun Microsystems (Solaris, Java), SuSE, Sybase, Symantec, Teradata (NCR), Trend Micro, Virage, webMethods, Wine, WMS Gaming, XyEnterprise, Yahoo!, and many others.

Tuesday, November 3, 2009

ICU is used many places

Where is ICU? - IBM

- Within IBM
  - All 5 major software brands
  - IBM operating systems
  - Ascential Software, Cognos, PSD Print Architecture, DB2, COBOL, Host Access Client, InfoPrint Manager, Informix GLS, iSeries, Language Analysis Systems, Lotus Notes, Lotus Extended Search, Lotus Workplace, WebSphere Message Broker, NUMA-Q, OTI, OmniFind, Pervasive Computing WECMS, Rational Business Developer and Rational Application Developer, SS&S Websphere Banking Solutions, Tivoli Presentation Services, Tivoli Identity Manager, WBI Adapter/ Connect/ Modeler and Monitor/ Solution Technology Development/WBI-Financial TePI, Websphere Application Server/ Studio Workload Simulator/ Transcoding Publisher, XML Parser…

ICU is used throughout IBM. It is also used by many other companies and organizations. Many of these companies and organizations also participate in improving ICU.

Where is ICU? - Google

- Within Google
  - Web Search
  - Chrome
  - Android
  - Adwords
  - Google Finance
  - Google Maps
  - Blogger
  - Google Analytics
  - Google Gears
  - Google Groups
  - others…

Agenda

- Isn't Unicode enough?
- Why ICU?
- Where is ICU?
- What's new in ICU?
- What's next for ICU?

What's New with ICU? — 4.2

- Timeline
  - September 1st, 2006: ICU 3.6
  - December 12th, 2007: ICU 3.8.1
  - July 2nd, 2008: ICU 4.0
  - **May 8th, 2009: ICU 4.2**
  - **Early 2010: ICU 4.4**

Tuesday, November 3, 2009

Where are we now? ICU 4.2 is the latest, and we are hard at work on 4.4

What's new with ICU? — CLDR 1.7 Locale Data

- 146 languages and 159 territories
- 468 locales in all
- 21% more locale data than previous release

Tuesday, November 3, 2009

ICU gets its locale data from the Unicode consortium's CLDR. ICU is also used as a test suite to verify CLDR's behavior.

What's new with ICU? — IETF BCP 47

- Replaces RFC 3066
- ICU:
  - ja_JP@calendar=japanese
- BCP47:
  - ja-JP-u-ca-japanese

What's new with ICU? — RBNF Rules from CLDR

- RBNF = Rule Based Number Format
  - "twelve thousand three hundred forty-five"
  - "一万二千三百四十五" (1×10000)+(2×1000)…
- Long-standing ICU feature
- Now part of and maintained by CLDR
- XML format rules

What's new with ICU? — Number Systems in Dates

- ICU could already change between decimal digit systems:
  - en: "2009", ar: "٢٠٠٩", hi: "२००९"
- Hebrew required different numbering systems within *one date*:
  - 13:41:34 ט"תשס ה'תשרי ח' (5769 8)

What's new with ICU? — Number Systems Overview

- CLDR 1.7 Feature
- Several CSS Numbering Systems supported
- Default system per locale
- Numbering systems can be simple (replacing 0-9) or complex (rule-based). All data in CLDR.
- Can specify different numbering systems for different parts of the date format.

What's new with ICU? — Encoding Selector

- Given an input string, returns the set of names of the corresponding converters which can convert the string
- "鳥суп" →
  - UTF-8
  - Shift_JIS
  - GBK
  - GB18030

What's new with ICU? — Simple Duration

- "3.5 years"
- "2 months"
- "0 mins", "1 min", "3 mins"

with appropriate plurals from CLDR

What's new with ICU? — StringPrep named profiles

- RFC3454 profiles available by keyword:
  - RFC3491 NAMEPREP, RFC3530 NFS4, RFC3722 iSCSI, RFC3920 NodePrep/ResourcePrep, RFC4011 MIB, RFC4013 SASLprep, RFC4505 trace and RFC4518 LDAPprep.

Tuesday, November 3, 2009

In previous versions, users had to build their own profile

What's new with ICU? — C++ API Ease of Use

- new UnicodeString methods
  - fromUTF32() & toUTF32()
  - fromUTF8() & toUTF8()
- UTF-8 as a fixed default codepage
- Collator API can compare UTF-8 strings directly

Tuesday, November 3, 2009

ICU4C Ease of Use
- new UnicodeString fromUTF32() & toUTF32() methods
- new UnicodeString fromUTF8() & toUTF8() methods
  no more preflighting with ustring.h functions to/from
    UnicodeString buffer, no more UConverter usage
  new StringPiece class for convenient UTF-8 input
    takes char* or std::string or other compatible string classes
  new ByteSink class for convenient UTF-8 output
    output to char[], std::string and other compatible string classes
- UTF-8 as fixed default codepage (ICU build option)
  set U_CHARSET_IS_UTF8 to 1
  makes functions taking char* arguments more efficient
  reduces size of statically linked ICU code
  enables efficient
    UnicodeString a("F\u00FC\u00DFe") // gcc
  or
    UnicodeString b("Füße") // .cpp in UTF-8

What's new with ICU? — C++ API Ease of Use

- C++ ErrorCode class simplifies UErrorCode handling
- easy conversion between UnicodeSet & USet

Tuesday, November 3, 2009

- C++ ErrorCode class simplifies UErrorCode handling
  does required initialization of UErrorCode (common source of bugs)
  easy syntax for use with both C & C++ APIs
  application overrides handleFailure() & destructor
  with error logging, throwing exception, etc.
  API change in ICU 4.4 (check() -> assertSuccess())
- easy conversion between UnicodeSet & USet
  simplify working with a mix of C and C++ APIs

What's new with ICU? — Miscellaneous

- Arabic shaping improvements
- Converter updates: ISCII (C+J), LMBCS (J)
- Query for CLDR Version
- UTS#39 spoof (confusable) detection (C)
- Locale Builder (J)
- BigDecimal MathContext for DecimalFormat (J)

What's improved in ICU?

- Performance/Footprint
  - Time Zone format/parse
  - DateIntervalFormat construction
  - Unicode data lookup
  - others
- Bug Fixes

Demo

Agenda

- Isn't Unicode enough?
- Why ICU?
- Where is ICU?
- What's new in ICU?
- What's next for ICU?

What's next for ICU?
(A tentative sampler)

- Early 2010: ICU 4.4
- Unicode 5.2, CLDR 1.8, BCP 47
- Performance and Usability
- Footprint improvement
- Lenient parsing improvements
- Java 5 migration including generics

References

- ICU main site:
  - http://icu-project.org
    - Download ICU
    - User Guide
    - Technical FAQ
    - Support
    - Bug Reports
    - Demonstrations

Tuesday, November 3, 2009

If you would like more information about ICU, you can go to our main site on http://icu-project.org There you will find links to download ICU, the ICU User Guide, the technical FAQ, where to get support, demonstrations of how ICU works, and many other topics related to ICU

You can also find more information about Unicode at the unicode.org site.