Locale-Aware Sorting and Text Handling in the Open Toolkit

美丽的话

Eliot Kimber
Contrext

DITA OT Day 2017

All translations via Google Translate—may be total nonsense

About the Author
关于作者

- Independent consultant focusing on DITA analysis, design, and implementation
- Doing SGML and XML for cough 30 years cough
- Founding member of the DITA Technical Committee
- Founding member of the XML Working Group
- Co-editor of HyTime standard (ISO/IEC 10744)
- Primary developer and founder of the DITA for Publishers project
- Author of *DITA for Practitioners*, Vol 1 (XML Press)

Requirements 要求

- Sort and group Simplified Chinese (and other languages)
- Do rough layout of lines
  - Requires finding work boundaries
  - Requires finding line-break opportunities
- Target applications
  - Sort and group glossaries, indexes, etc.
  - Fixed-layout EPUBs
- Free, open-source solution

Challenge: Simplified Chinese
挑战: 简体中文

- Simplified Chinese collates using pinyin transliteration of ***words***
- All other languages collate based on individual ***characters***
  - Alphabetic and syllabic languages
  - Traditional Chinese: Radical and stroke count
  - Can be done with static per-character configuration

Simplified Chinese Requires a Dictionary
简体中文需要一个词典

- Pinyin transliteration depends on whole word
  - Same character may have different transliteration in different words (多音字)
  - As if “c” was pronounced “see” by itself but “tee” in “cat”
  - E.g.: 行 – xíng or hāng
- Requires a dictionary with pinyin transliteration
  - Requires ability to find words
- Need word frequency as secondary sort parameter or to disambiguate words
  - Requires frequency dictionary

Solution 解

- Use CC-CEDICT open-source dictionary
  - ~155,000.00 entries
- ICU4J for word and line breaking
- Java 2D for text length approximation
- ICU4J for grouping and sorting of other languages
- Integrated with Saxon's collator Java API
- XSLT functions for use from XSLT
- Configured as DITA OT plugin for use in OT
- Java and XSLT library not OT-specific

Result: DITA Community i18n Plugin
结果：DITA社区i18n插件

- Java: Custom collator for Simplified Chinese
  - Implements Saxon RuleBasedCollator
  - Can be used with Saxon 9.1+ or any Java code
  - Locale-aware text analysis and metrics
- XSLT function library
  - Grouping and sorting keys
  - Word and line breaking
  - Rendered text length given font details
- General grouping and sorting configuration file mechanism

Under Active Development
积极发展

- More accurate pinyin determination
- Use of word frequency data as secondary sort key

Not Yet Implemented
尚未落实

- Grouping for glossaries
  - And other things you might want to group
- Extension/replacement for built-in OT index grouping and sorting

Demo 演示

If time permits

Questions? 问题

Resources 资源

- Me: ekimber@contrext.com
- DITA Community i18n project: https://github.com/dita-community/org.dita-community.i18n
- http://blog.tutorming.com/mandarin-chinese-learning-tips/chinese-characters-with-various-pronunciations