Unicode in Action
Cummings, McKenna, Texin

Internationalization and Unicode Conference 40

November 1, 2016
Presenters

• Tex Texin
  Globalization Architect, XenCraft

• Craig R. Cummings
  Staff Consultant/Globalization Evangelist, VMware

• Michael McKenna
  I18n Product Owner, PayPal, Inc.
• The demo code will be available on I18nGuy.com shortly after the conference
The Unicode in Action tutorial is a 90 minute session that demonstrates programming with Unicode and related best practices.

This tutorial will build a simple application and demonstrate the code and resulting behavior as internationalization functions are added. Attendees will be able to relate these prototype examples to the requirements of their own applications and reference them to code solutions.

The program will show sorting of different strengths, regular expressions, Unicode normalization, bidirectional languages, and other features of the Unicode standard. The tutorial will highlight why each of these functions are needed so you can determine when to use them in your applications.
Objectives

• Be introductory level
• Simple examples
• The program will show
  – sorting of different strengths,
  – regular expressions,
  – Unicode normalization,
  – bidirectional languages,
  – and other features.
  – Highlight the need for these features.
Base Program – Movie Catalog

• Our first example is a simple movie catalog.
• It could be any business application, listing products, customers, etc.
• It demonstrates typical data requirements:
  – text, dates, numbers, currencies, taxonomies, images.
• It is written in HTML5 and JavaScript
  – For simplicity and availability
  – Turns out, not all that portable. Firefox for now
## Base Program – Movie Catalog

### Movie Catalog

<table>
<thead>
<tr>
<th>Title</th>
<th>Release Date</th>
<th>Genre</th>
<th>Units (Thousands)</th>
<th>Price</th>
<th>Cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast &amp; Furious</td>
<td>1/21/2001</td>
<td>Action</td>
<td>8,106</td>
<td>$3.50</td>
<td><img src="image1.png" alt="Image" /></td>
</tr>
<tr>
<td>Hackers</td>
<td>1/21/1995</td>
<td>Crime</td>
<td>10,709.67</td>
<td>$2.65</td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td>Jurassic Park</td>
<td>1/21/1993</td>
<td>Sci-Fi</td>
<td>1,275.8</td>
<td>$6,543.21</td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
<tr>
<td>Shocker</td>
<td>1/21/1989</td>
<td>Horror</td>
<td>90,109</td>
<td>$4.12</td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Unicode in Action Movie Catalog

```
<!DOCTYPE html>
<html>
<head>
  <meta charset="utf-8">
  <title>Unicode in Action Movie Catalog</title>
  <link href="css/styles.css" rel="stylesheet">
</head>
<body>

...<h1>Options</h1>
<form id="options" name="settings" onsubmit="return myControls();" >
  <p>Search: <input type="text" name="search" size="40"
    placeholder="search term or regular expression"></p>

  <div class="controlbuttons">
    <input type="submit" value="Go">
  </div>
</form>

<div id="datalist">
  <table class='products-list'>
    <caption>Movie Catalog</caption>
    <tr id="prodheading">
      <th>Title</th>
      <th>Release Date</th>
      <th>Genre</th>
      <th>Units<br>(Thousands)</th>
      <th>Price</th>
      <th>Cover</th>
    </tr>
    <tbody id="id01">
      <tr>
        <td>Go Go</td>
        <td>2018-01-01</td>
        <td>Action</td>
        <td>12,345</td>
        <td>9.99</td>
        <td>Poster 1</td>
      </tr>
    </tbody>
  </table>
</div>
```

Unicode in Action
<script type="text/javascript">
    function getProducts() {
        var products = readjson("","products.json");
        showProducts(products);
    }

    function myControls() {
        UIApattern =
        document.forms["settings"]["search"].value;

        getProducts();
        return false;
    }

    function myimage(value) {
        var intlvalue = "<img alt='movie cover photo' src=" + value + ">
        return(intlvalue);
    }
</script>

/* return true for records that do not match*/
function searchFilter(testValue, matchPattern) {
    var exclude = false;
    if (matchPattern == ") { return (exclude);
    }
    var REpattern = new RegExp(matchPattern, "i");
    exclude = (testValue.search(REpattern) == -1);
    /* if not found, exclude = true */
    return (exclude);
}
function showProducts(data) {
    var i;
    var out = "";
    for(i = 0; i < data.length; i++) {
        if (searchFilter(data[i].title, UIApattern)) {
            continue;
        }
        out += "<tr><td>" + data[i].title + "</td><td>" + mydate(data[i].specs.year) + "</td><td>" + mygenre(data[i].specs.genre) + "</td><td>" + mynumber(data[i].specs.duration) + "</td><td>" + mycurrency(data[i].price) + "</td><td>" + myimage(data[i].image.small) + "</td></tr>
    }

    document.getElementById("id01").innerHTML = out;
}
What do we need to make this program global?
## Base Program – Movie Catalog

### Options

Search: search term or regular expression

- **Locale**, **Search**, **Sort**, **Normalization**, **Bidi**, **LTR**, **RTL**, **Encoding** (UTF-8, UTF-16, Supplementary Characters)

### Movie Catalog

<table>
<thead>
<tr>
<th>Title</th>
<th>Release Date</th>
<th>Genre</th>
<th>Units (Thousands)</th>
<th>Price</th>
<th>Cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast &amp; Furious</td>
<td>1/21/2001</td>
<td>Action</td>
<td>8,106</td>
<td>$3.50</td>
<td></td>
</tr>
<tr>
<td>Hackers</td>
<td>1/21/1995</td>
<td>Crime</td>
<td>10,709.67</td>
<td>$2.65</td>
<td></td>
</tr>
<tr>
<td>Jurassic Park</td>
<td>1/21/1993</td>
<td>Sci-Fi</td>
<td>1,275.8</td>
<td>$6,543.21</td>
<td></td>
</tr>
<tr>
<td>Shocker</td>
<td>1/21/1989</td>
<td>Horror</td>
<td>90,109</td>
<td>$4.12</td>
<td></td>
</tr>
</tbody>
</table>
### Internationalized Movie Catalog

**Chinese Locale**

---

**Unicode in Action**

**MOVIE CATALOG**

#### Options

- **Search**: search term or regular expression
- **Sort Direction**: Ascending, Descending
- **Sort Strength**: Accent
- **Normalization**: NFC, NFKC, NFD, NFKD
- **Layout Direction**: LTR, RTL
- **Locale**: Chinese

---

#### Movie Catalog

<table>
<thead>
<tr>
<th>标题</th>
<th>发布日期</th>
<th>类型</th>
<th>单元（干）</th>
<th>价格</th>
<th>封面</th>
</tr>
</thead>
<tbody>
<tr>
<td>„The Walk“ bringt Zuschauer zum Erbrechen</td>
<td>2015/1/21</td>
<td>外国</td>
<td>565,674.99</td>
<td>¥ 4.70</td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>A Royal Night - Ein königliches Vergnügen</td>
<td>2015/4/21</td>
<td>外国</td>
<td>9,876.54</td>
<td>¥ 4.23</td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>Arbeit macht das Leben süß, Faulheit stärkt die Glieder</td>
<td>2015/4/21</td>
<td>外国</td>
<td>1,246.89</td>
<td>¥ 4.23</td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>Die Kleinen und die Bösen</td>
<td>2005/4/11</td>
<td>外国</td>
<td>8,643.21</td>
<td>¥ 4.23</td>
<td><img src="image" alt="Image" /></td>
</tr>
</tbody>
</table>
## Internationalized Movie Catalog

### Arabic Locale, RTL Direction

<table>
<thead>
<tr>
<th>Movie Catalog</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cover</strong></td>
</tr>
<tr>
<td><img src="image1" alt="Cover" /></td>
</tr>
<tr>
<td><img src="image2" alt="Cover" /></td>
</tr>
<tr>
<td><img src="image3" alt="Cover" /></td>
</tr>
<tr>
<td><img src="image4" alt="Cover" /></td>
</tr>
<tr>
<td><img src="image5" alt="Cover" /></td>
</tr>
<tr>
<td><img src="image6" alt="Cover" /></td>
</tr>
<tr>
<td><img src="image7" alt="Cover" /></td>
</tr>
</tbody>
</table>

### Options

- **Search**
- **Sort Direction**
  - Ascending
  - Descending
- **Sort Strength**
  - Accent
- **Normalization**
  - NFC
  - NFKC
  - NFD
  - NFkd
- **Layout Direction**
  - LTR
  - RTL
- **Locale**
  - Arabic

### Unicode in Action

- MOVIE CATALOG
Internationalized Movie Catalog Features

• Uses locales
  – (en-US, de-DE, zh-CN, sv, ar)
• Localized headings, taxonomy
• Formatted data (date, number, price)
• Normalization of input
• Localized sort
• Bidi
Normalization

Tex Texin
Internationalization Architect
Canonical & Compatibility Normalization

• Unicode characters can have more than 1 representation

• Canonical equivalence
  – Indistinguishable, fundamental equivalence
  – E.g. combining sequences, singletons
  – “Å” U+00C5 (A-ring pre-composed)
  – “A+˚ ” U+0041 + U+030A (A + combining ring above)
  – “Å” U+212B (Angstrom)

• Compatibility equivalence
  – E.g. Formatting differences, ligatures
  – “カ” U+FF76 “カ” U+30AB (KA half and full width)
  – “fi” U+FB01 (ligature fi)
Unicode Consortium has defined canonical and compatibility decomposition formats and 4 different sets of rules for normalization:

“Unicode Normalization Forms”

http://www.unicode.org/unicode/reports/tr15/

<table>
<thead>
<tr>
<th></th>
<th>Composed</th>
<th>Decomposed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canonical</strong></td>
<td>NFC</td>
<td>NFD</td>
</tr>
<tr>
<td><strong>Canonical+ Kompatibility</strong></td>
<td>NFKC</td>
<td>NFKD</td>
</tr>
</tbody>
</table>
Sorting

Tex Texin
Internationalization Architect
Collation

Dependencies

• Language
• Application
  – Dictionary
  – Phonebook
• “Strength”
  – Accent
  – Case
  – Ignorables
## Example Collation Differences

<table>
<thead>
<tr>
<th>Language</th>
<th>Swedish:</th>
<th>z &lt; ö</th>
</tr>
</thead>
<tbody>
<tr>
<td>German:</td>
<td>ö &lt; z</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usage</th>
<th>Dictionary:</th>
<th>öf &lt; of</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Telephone:</td>
<td>of &lt; öf</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Customizations</th>
<th>Upper–first</th>
<th>A &lt; a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower–First</td>
<td>a &lt; A</td>
</tr>
</tbody>
</table>
### Comparison Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>Base characters</td>
<td>role &lt; roles &lt; rule</td>
</tr>
<tr>
<td>L2</td>
<td>Accents</td>
<td>role &lt; rôle &lt; roles</td>
</tr>
<tr>
<td>L3</td>
<td>Case</td>
<td>role &lt; Role &lt; rôle</td>
</tr>
<tr>
<td>L4</td>
<td>Punctuation</td>
<td>role &lt; “role” &lt; Role</td>
</tr>
<tr>
<td>Ln</td>
<td>Tie–Breaker</td>
<td>role &lt; roôle &lt; “role”</td>
</tr>
</tbody>
</table>

Box represents format character

Purple chars more significant than differences indicated by underscores
### Accent Ordering

<table>
<thead>
<tr>
<th>Forward Accent Ordering</th>
<th>cote &lt; coté &lt; côte &lt; côté</th>
</tr>
</thead>
<tbody>
<tr>
<td>French Accent Ordering</td>
<td>cote &lt; côte &lt; coté &lt; côté</td>
</tr>
</tbody>
</table>

French gives more weight to accents at the end of the string than the beginning. Cote and Coté are more similar in forward ordering, but in French, Côte orders between the two.
Language Identifiers

Tex Texin
Internationalization Architect

Unicode in Action
Language Identification

- **HTTP:** Content-Language header
- **HTML:** LANG attribute
e.g. `<html lang="fr">`
- **XML:** xml:lang attribute
- **XHTML 1.0:** Both lang and xml:lang
  `<p xml:lang="la" lang="la">Verba.</p>`
- **XHTML 1.1, 2:** xml:lang attribute
### BCP47 Language Identifiers

<table>
<thead>
<tr>
<th>Subtag</th>
<th>Standard</th>
<th>Syntax</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>ISO 639</td>
<td>2 or 3 letter code</td>
<td>en, yue</td>
</tr>
<tr>
<td>Extlang</td>
<td>ISO 639-2</td>
<td>3 letter code</td>
<td>(Legacy only) zh-yue</td>
</tr>
<tr>
<td>Script</td>
<td>ISO 15924</td>
<td>4 letter code</td>
<td>Latn, Cyrl, Hans, Hant</td>
</tr>
<tr>
<td>Region</td>
<td>ISO 3166</td>
<td>2 letter code</td>
<td>US, GB 419</td>
</tr>
<tr>
<td></td>
<td>UN M49</td>
<td>3 digit code</td>
<td></td>
</tr>
<tr>
<td>variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>extensions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>privateuse</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tag</th>
<th>Language</th>
<th>Tag</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>English</td>
<td>zh</td>
<td>Chinese</td>
</tr>
<tr>
<td>en-US</td>
<td>American English</td>
<td>zh-Hant</td>
<td>Traditional Chinese</td>
</tr>
<tr>
<td>es-US</td>
<td>Spanish as spoken in U.S.</td>
<td>zh-Hans</td>
<td>Simplified Chinese</td>
</tr>
<tr>
<td>en-CA</td>
<td>Canadian English</td>
<td>cmn</td>
<td>Mandarin</td>
</tr>
<tr>
<td>fr-CA</td>
<td>Canadian French</td>
<td>yue</td>
<td>Cantonese</td>
</tr>
<tr>
<td>es-ES</td>
<td>Iberian Spanish</td>
<td>cmn-Hant</td>
<td>Mandarin in Traditional Chinese</td>
</tr>
<tr>
<td>es-419</td>
<td>Latin American Spanish</td>
<td>pt-BR</td>
<td>Brazilian Portuguese</td>
</tr>
<tr>
<td>es-MX</td>
<td>Mexican Spanish</td>
<td>zh-yue</td>
<td>retired, use yue instead</td>
</tr>
<tr>
<td></td>
<td></td>
<td>zh-CN</td>
<td>Chinese spoken in China</td>
</tr>
</tbody>
</table>
Language Identification – CSS

Two methods use the language attribute in CSS:

• The `lang` pseudo-class.

```css
*:lang(zh) { font-family: SimSun }
```

• The attribute selector.

```css
*[lang|=fr] { font-weight:bold }
```

• Both use the same matching mechanism as the `lang()` function in XPath.

➤ Example: [LanguagesCSS.htm](#)
# Text Layout Standards

More content and example code are available at:

[www.xencraft.com/training/webstandards.html](http://www.xencraft.com/training/webstandards.html)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lang()</td>
<td>Xsl format-number</td>
</tr>
<tr>
<td>Lang pseudo-class</td>
<td>Html bi-directional text</td>
</tr>
<tr>
<td>Lang attr selector</td>
<td>Css bi-directional text</td>
</tr>
<tr>
<td>Quote:qo</td>
<td>Vertical text (SVG losing ground)</td>
</tr>
<tr>
<td>Text-transform</td>
<td>Ruby annotation</td>
</tr>
<tr>
<td>Css list-style-type</td>
<td>Css3 combined sort</td>
</tr>
<tr>
<td>Xsl number</td>
<td>Xsl:sort</td>
</tr>
</tbody>
</table>
Bidirectional Support

Tex Texin
Internationalization Architect
Bidirectional (Bidi) Language Support

• **HTML 4 DIR attribute**
  
  ```html
  dir="ltr" | dir="rtl"
  ```
  
  – Sets base direction
  
  – Direction is inherited

• **Direction affects alignment and flow**

  – Ordering of text and table columns
  
  – Text alignment, Alignment of overflowing blocks

• **Control Characters**

  – Right to Left and Left to Right Marks `&rlm;`/`&lrm;`
  
  – Useful for correct positioning of neutrals
Bidirectional (Bidi) Language Support

- HTML 5 – Isolates
  
  `<bdi dir=rtl> </bdi>`

- Flow doesn’t change with container changes!

- DIR=AUTO
  - Detects direction, based on first strong character

- CSS Selectors
  - :dir (rtl) for rtl elements
  - :dir(ltr) for ltr elements
## Internationalized Movie Catalog

### Arabic Locale, RTL Direction

### Movie Catalog

<table>
<thead>
<tr>
<th>Cover</th>
<th>Price</th>
<th>Units (Thousands)</th>
<th>Genre</th>
<th>Release Date</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Cover" /></td>
<td>₹3.69</td>
<td>2,46,189</td>
<td>العربية</td>
<td>3-10-81</td>
<td>The Walk: Zuschauer zum Erbrechen (41)</td>
</tr>
<tr>
<td><img src="image2" alt="Cover" /></td>
<td>₹9,34,189</td>
<td></td>
<td>العربية</td>
<td>3-10-81</td>
<td>A Royal Night: Ein königliches Vergnügen (41)</td>
</tr>
<tr>
<td><img src="image3" alt="Cover" /></td>
<td>₹1,743,19</td>
<td></td>
<td>العربية</td>
<td>3-10-81</td>
<td>Arbeit macht das Leben süß, Faulheit stärkt die Glieder (55)</td>
</tr>
<tr>
<td><img src="image4" alt="Cover" /></td>
<td>₹0.59</td>
<td>8,136,11</td>
<td>العربية</td>
<td>3-10-81</td>
<td>Die Kleinen und die Bösen (25)</td>
</tr>
<tr>
<td><img src="image5" alt="Cover" /></td>
<td>₹3.39</td>
<td>18,789</td>
<td>العربية</td>
<td>3-10-81</td>
<td>Die Legende der weißen Pferde (29)</td>
</tr>
<tr>
<td><img src="image6" alt="Cover" /></td>
<td>₹8,51</td>
<td>1,18,76</td>
<td>العربية</td>
<td>3-10-81</td>
<td>Fast &amp; Furious (14)</td>
</tr>
<tr>
<td><img src="image7" alt="Cover" /></td>
<td>₹6,51</td>
<td>3,718,6</td>
<td>العربية</td>
<td>3-10-81</td>
<td>Gotthard Graubner - Farb-Raum-Körper (36)</td>
</tr>
</tbody>
</table>

### Options

- **Search**: input search term or regular expression
- **Sort Direction**: Ascending, Descending
- **Sort Strength**: Accent, Normalization
- **Normalization**: NFC, NFKC, NFD, NFKD, None
- **Layout Direction**: LTR, RTL
- **Locale**: Arabic

---

Unicode in Action

33
Bidi References

- W3C Bidi Tutorial
  - www.w3.org/International/tutorials/bidi-xhtml/
- Inline markup and bidirectional text in HTML
  - www.w3.org/International/articles/inline-bidi-markup/
- Additional Requirements for Bidi in HTML and CSS
  - www.w3.org/TR/html-bidi/
- Unicode Bidirectional Algorithm
  - (Unicode Standard Annex #9)
  - www.unicode.org/reports/tr9/
- A Tale of Opposing Directions: Bidirectional Text in HTML and CSS
  - Elika J. Etemad (fantasai) Mozilla Project W3C CSS Working Group
  - fantasai.inkedblade.net/style/talks/bidi/
Character Counting

Tex Texin
Internationalization Architect
• How long is a string?
  – Ångstrom 8 or 9 characters?
  – fire 3 or 4 characters?
  – 😱 1 or 2 characters?
Character Counting, Indexing, Length

• How long is a string?
  – Ångstrom 8 or 9 characters
    • 8 if composed characters, 9 if combining
    • A + combining ring above U+030A
  – fire 3 or 4 characters
    • 3 if “fi” is a ligature U+FB01

  – 😱 1 or 2 characters?
    • 1 if an abstract character 0x1F631
    • 2 if UTF-16 code units \uD83D\uDE31
    • www.i18nguy.com/unicode/surrogatetable.html
• Beware inconsistencies in your code as well as your platforms
  – JavaScript supports 6 digit escapes 0x1F631
  – JSON uses surrogates \uD83D\uDE31
  – string.length counts abstract characters
  – string.substring counts code points
  – Both treat combining characters as separate characters
  – Both treat a ligature as one character
    • Normalization can aid consistency
Internationalized Code

Tex Texin
Internationalization Architect
<!DOCTYPE html>
<html lang="en" dir="ltr" id="html01">
<head>
  <meta charset="utf-8">
</head>
function collSort(data, locale, sortDir, strength) {
    var coll = Intl.Collator(locale, {sensitivity : strength});

    for (var i = data.length - 1; i >= 0; i--) {

        for (var j = 0; j < i; j++) {
            if (sortDir == "asc") {
                if (coll.compare(data[j].title, data[j+1].title) > 0) _swap(data, j, j+1);
            } else {
                if (coll.compare(data[j].title, data[j+1].title) < 0) _swap(data, j, j+1);
            }
        }
    }
}
function mydate (value) {
    var datevalue= new Date(value);
    var intlvalue = new Intl.DateTimeFormat(UIAlocale).format(datevalue);
    return(intlvalue);
}

function mynumber (value) {
    var intlvalue = new Intl.NumberFormat(UIAlocale).format(value);
    return(intlvalue);
}
function mycurrency (value) {
    var currencylist =
        {'en-US': 'USD', 'de-DE':'EUR', 'zh-CN': 'CNY', 'ar': 'SAR', 'sv': 'SEK'};

    var mycur = currencylist[UIAlocale];

    var intlvalue = new Intl.NumberFormat(UIAlocale, {
        style: 'currency', currency: mycur
    }).format(value);
    return(intlvalue);
}
Normalization of Input

If the selected normalization form is “none” do not normalize the search string. Otherwise, normalize the string into the selected form. One of: NFC, NFKC, NFD, NFKD

```javascript
function mySearch(NormForm, SearchString) {
    if (NormForm == "none") {
        return (SearchString); /* nothing to do */
    }
    return (SearchString.normalize(NormForm));
}
```
Tex Texin

Tex is an industry thought leader specializing in business and software globalization services. His expertise includes global product strategy, Unicode and internationalization architecture, and cost-effective implementation and testing. Over the past two decades, Tex has created numerous global products, led internationalization development teams, and guided companies in taking business to new regional markets.

Tex is a contributor to internationalization standards for software and on the Web.

Tex is a popular speaker at conferences around the world and provides on-site training on Unicode, internationalization, and globalization QA worldwide.

Tex is the author of the popular, instructional web site www.I18nGuy.com

Tex is founder and Chief Globalization Architect for XenCraft. XenCraft provides global business consulting and software design, implementation, test and training services on globalization product strategy and software internationalization architecture.

Unicode in Action