Unicode in Action

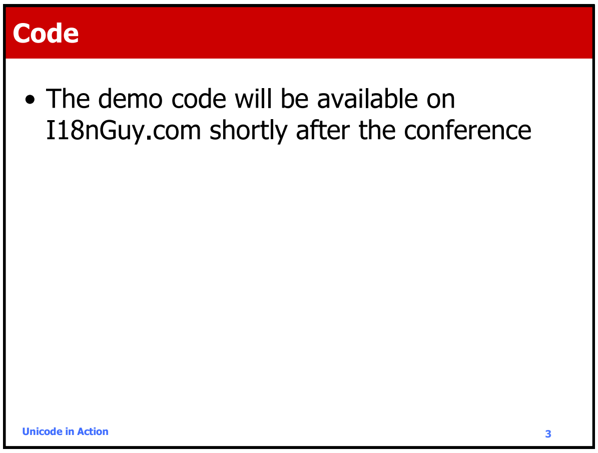
Cummings, McKenna, Texin

Internationalization and Unicode Conference 42

Sept. 10, 2018

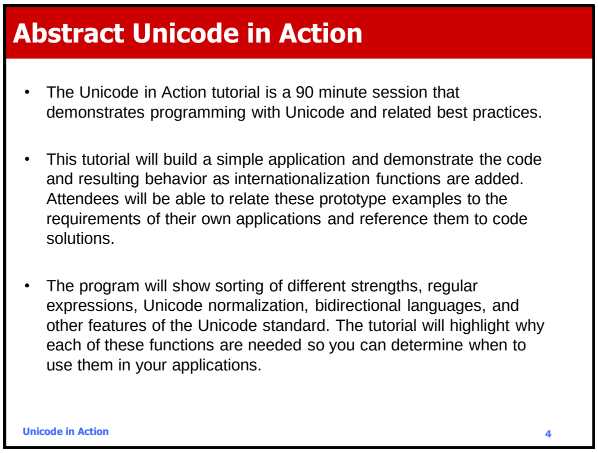
Presenters

- Tex Texin
  Globalization Architect, XenCraft
- Craig R. Cummings
  Staff Consultant/Globalization Evangelist, VMware
- Michael McKenna
  World-Ready Architect, PayPal, Inc.



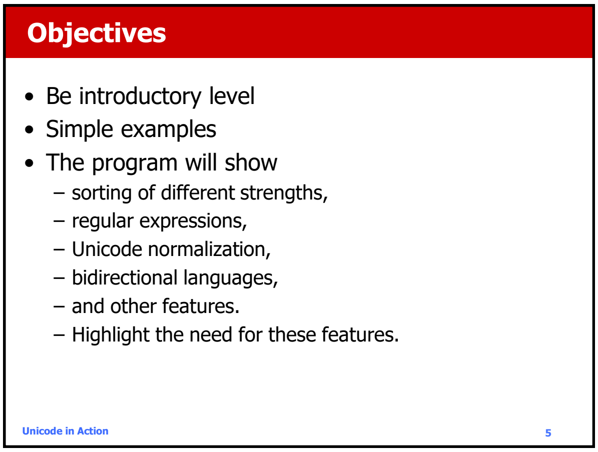
Code

- The demo code will be available on I18nGuy.com shortly after the conference



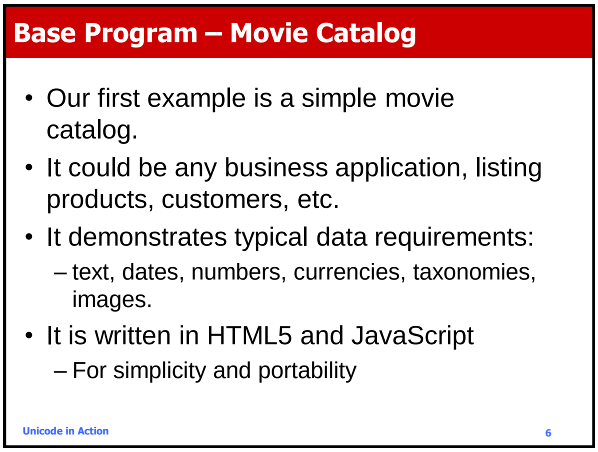
Abstract Unicode in Action

- 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.



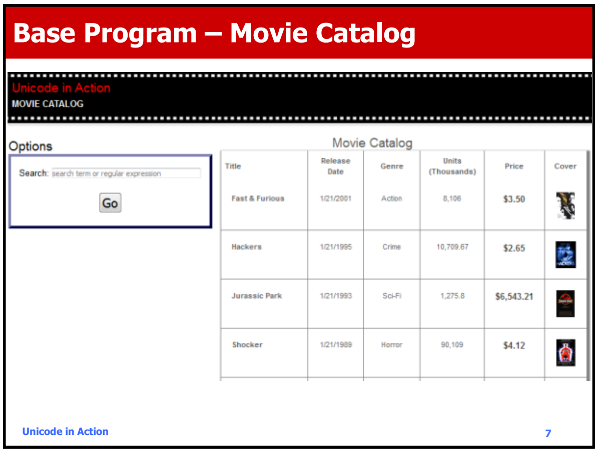
Objectives

- 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 portability



Base Program – Movie Catalog

Unicode in Action
MOVIE CATALOG

Options

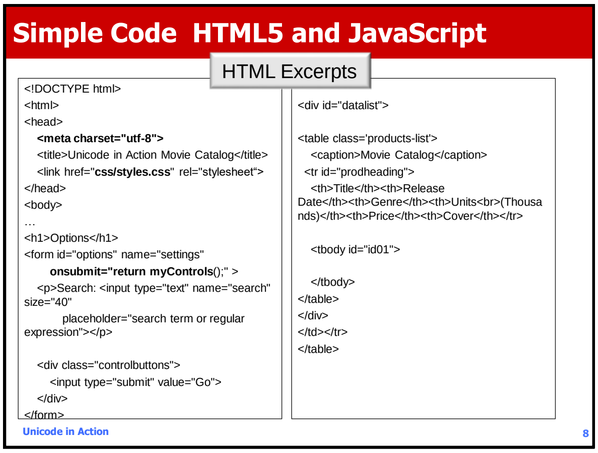
Search: search term or regular expression

Go

Movie Catalog

Title	Release Date	Genre	Units (Thousands)	Price	Cover
Fast & Furious	1/21/2001	Action	8,106	$3.50	
Hackers	1/21/1995	Crime	10,709.67	$2.65	
Jurassic Park	1/21/1993	Sci-Fi	1,275.8	$6,543.21	
Shocker	1/21/1989	Horror	90,109	$4.12	

Unicode in Action 7



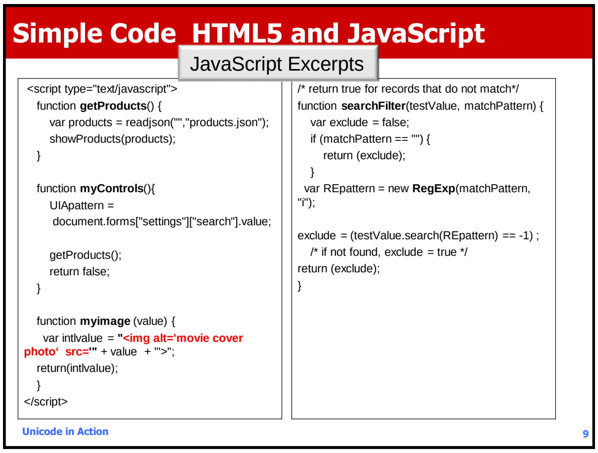
Simple Code HTML5 and JavaScript

HTML Excerpts

```
<!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>(Thousa
nds)</th><th>Price</th><th>Cover</th></tr>

   <tbody id="id01">

   </tbody>
</table>
</div>
</td></tr>
</table>
```

Unicode in Action 8



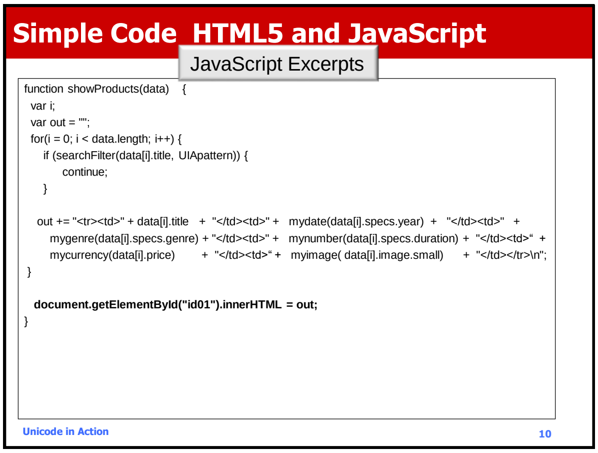
Simple Code HTML5 and JavaScript

JavaScript Excerpts

```
 <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);
}
```

Unicode in Action 9

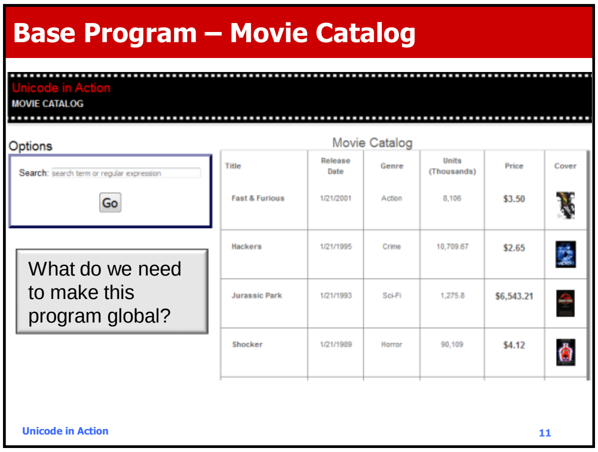


Simple Code HTML5 and JavaScript

JavaScript Excerpts

```
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>\n";
}

 document.getElementById("id01").innerHTML = out;
}
```

Unicode in Action 10



Base Program – Movie Catalog

Unicode in Action
MOVIE CATALOG

Options

Search: search term or regular expression

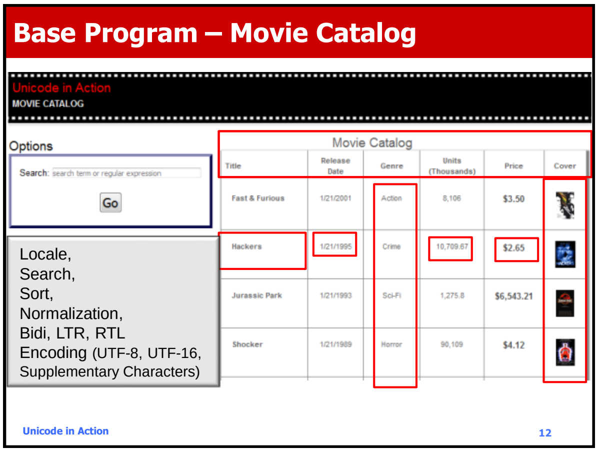
Go

Movie Catalog

Title	Release Date	Genre	Units (Thousands)	Price	Cover
Fast & Furious	1/21/2001	Action	8,106	$3.50	
Hackers	1/21/1995	Crime	10,709.67	$2.65	
Jurassic Park	1/21/1993	Sci-Fi	1,275.8	$6,543.21	
Shocker	1/21/1989	Horror	90,109	$4.12	

What do we need to make this program global?

Unicode in Action 11



Base Program – Movie Catalog

Unicode in Action
MOVIE CATALOG

Options

Search: search term or regular expression

Go

Movie Catalog

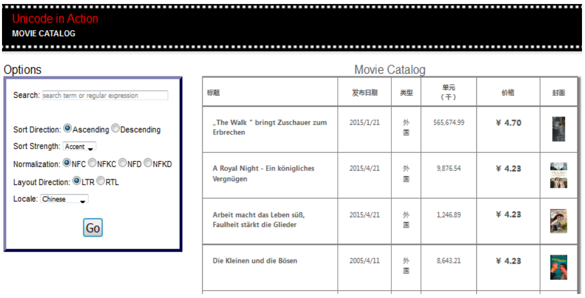
Title	Release Date	Genre	Units (Thousands)	Price	Cover
Fast & Furious	1/21/2001	Action	8,106	$3.50	
Hackers	1/21/1995	Crime	10,709.67	$2.65	
Jurassic Park	1/21/1993	Sci-Fi	1,275.8	$6,543.21	
Shocker	1/21/1989	Horror	90,109	$4.12	

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

Unicode in Action 12

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

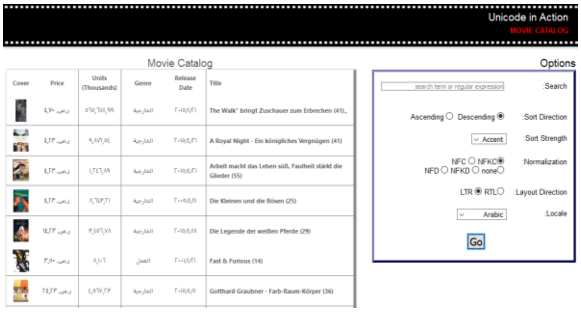
Go

Movie Catalog

标题	发布日期	类型	单元（千）	价格	封面
„The Walk " bringt Zuschauer zum Erbrechen	2015/1/21	外国	565,674.99	¥ 4.70	
A Royal Night - Ein königliches Vergnügen	2015/4/21	外国	9,876.54	¥ 4.23	
Arbeit macht das Leben süß, Faulheit stärkt die Glieder	2015/4/21	外国	1,246.89	¥ 4.23	
Die Kleinen und die Bösen	2005/4/11	外国	8,643.21	¥ 4.23	

Internationalized Movie Catalog

Arabic Locale, RTL Direction



Unicode in Action
MOVIE CATALOG

Movie Catalog

Cover	Price	Units (Thousands)	Genre	Release Date	Title
	٤٫٧٠ ر.س.	٥٦٥٬٦٧٤٫٩٩	الخارجية	٢٠١٥/١/٢١	The Walk" bringt Zuschauer zum Erbrechen (41)„
	٤٫٢٣ ر.س.	٩٬٨٧٦٫٥٤	الخارجية	٢٠١٥/٤/٢١	A Royal Night - Ein königliches Vergnügen (41)
	٤٫٢٣ ر.س.	١٬٢٤٦٫٨٩	الخارجية	٢٠١٥/٤/٢١	Arbeit macht das Leben süß, Faulheit stärkt die Glieder (55)
	٤٫٢٣ ر.س.	٨٬٦٤٣٫٢١	الخارجية	٢٠٠٥/٤/١١	Die Kleinen und die Bösen (25)
	١٤٫٢٣ ر.س.	٣٬٤٥٦٫٧٨	الخارجية	٢٠١٥/٤/١٨	Die Legende der weißen Pferde (29)
	٣٫٥٠ ر.س.	٨٬١٠٦	العمل	٢٠٠١/٦/٢١	Fast & Furious (14)
	٢٤٫٢٣ ر.س.	٤٬٥٦٧٫٢٣	الخارجية	٢٠١٥/٤/٨	Gotthard Graubner - Farb-Raum-Körper (36)

Options

search term or regular expression :Search

Ascending ○ Descending ● :Sort Direction

Accent :Sort Strength

NFC ○ NFKC● NFD ○ NFKD ○ none○ :Normalization

LTR ● RTL○ :Layout Direction

Arabic :Locale

Go

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

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 Normalization Forms

- 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/

	Composed	Decomposed
Canonical	NFC	NFD
Canonical+ Kompatibility	NFKC	NFKD

Sorting

Tex Texin
Internationalization Architect

Collation

Dependencies

- Language
- Application
  - Dictionary
  - Phonebook
- “Strength”
  - Accent
  - Case
  - Ignorables

Example Collation Differences

Language	Swedish:	z < ö
	German:	ö < z
Usage	Dictionary:	öf < of
	Telephone:	of < öf
Customizations	Upper-first	A < a
	Lower-First	a < A

Comparison Levels

Level	Description	Examples
L1	Base characters	role < roles < rule
L2	Accents	role < rôle < roles
L3	Case	role < Role < rôle
L4	Punctuation	role < “role” < Role
Ln	Tie-Breaker	role < ro□le < “role”

Box represents format character

Purple chars more significant than differences indicated by underscores

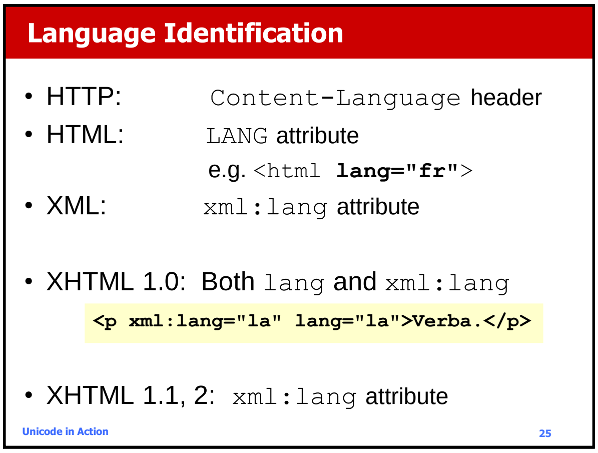
Accent Ordering

Forward Accent Ordering	cote < coté < côte < côté
French Accent Ordering	cote < côte < coté < côté

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



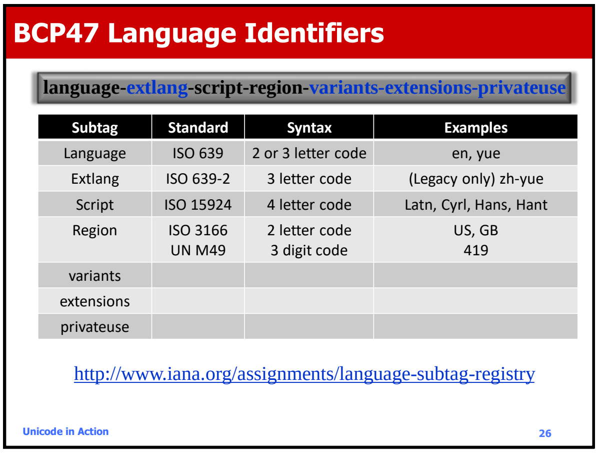
Language Identification

- 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

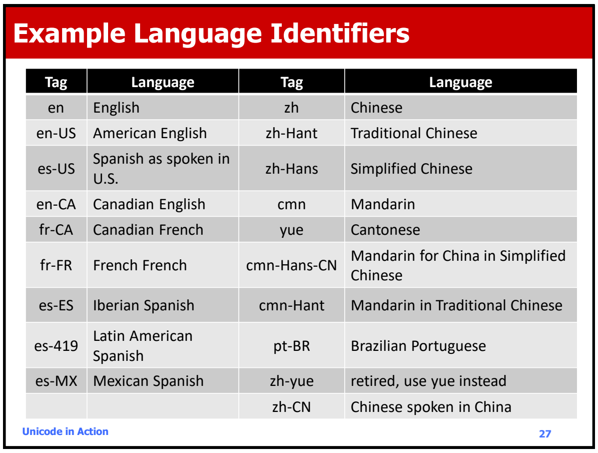


BCP47 Language Identifiers

language-extlang-script-region-variants-extensions-privateuse

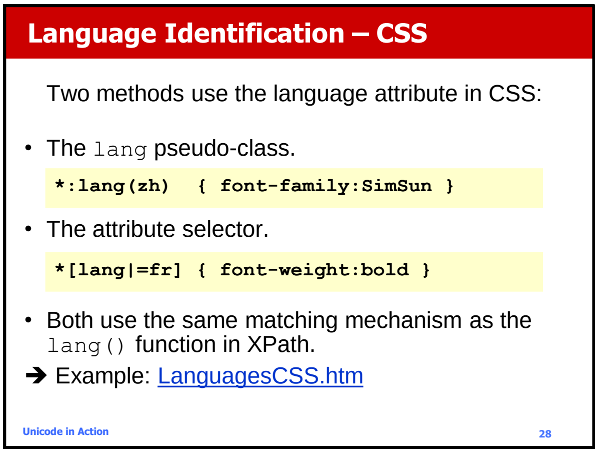
Subtag	Standard	Syntax	Examples
Language	ISO 639	2 or 3 letter code	en, yue
Extlang	ISO 639-2	3 letter code	(Legacy only) zh-yue
Script	ISO 15924	4 letter code	Latn, Cyrl, Hans, Hant
Region	ISO 3166 UN M49	2 letter code 3 digit code	US, GB 419
variants			
extensions			
privateuse			

http://www.iana.org/assignments/language-subtag-registry



Example Language Identifiers

Tag	Language	Tag	Language
en	English	zh	Chinese
en-US	American English	zh-Hant	Traditional Chinese
es-US	Spanish as spoken in U.S.	zh-Hans	Simplified Chinese
en-CA	Canadian English	cmn	Mandarin
fr-CA	Canadian French	yue	Cantonese
fr-FR	French French	cmn-Hans-CN	Mandarin for China in Simplified Chinese
es-ES	Iberian Spanish	cmn-Hant	Mandarin in Traditional Chinese
es-419	Latin American Spanish	pt-BR	Brazilian Portuguese
es-MX	Mexican Spanish	zh-yue	retired, use yue instead
		zh-CN	Chinese spoken in China



Language Identification – CSS

Two methods use the language attribute in CSS:

- The `lang` pseudo-class.

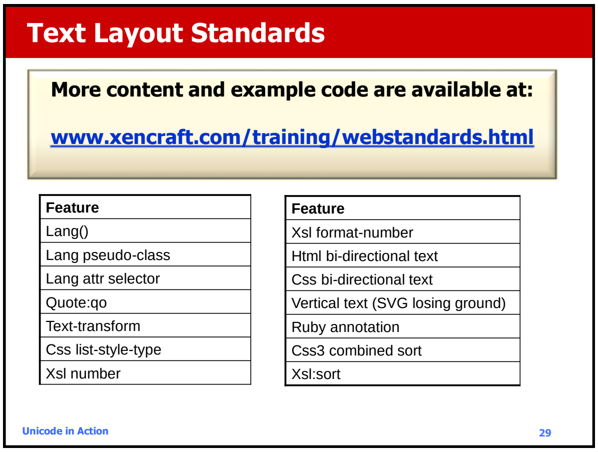
```
*:lang(zh)  { font-family:SimSun }
```

- The attribute selector.

```
*[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

Feature
Lang()
Lang pseudo-class
Lang attr selector
Quote:qo
Text-transform
Css list-style-type
Xsl number

Feature
Xsl format-number
Html bi-directional text
Css bi-directional text
Vertical text (SVG losing ground)
Ruby annotation
Css3 combined sort
Xsl:sort

Bidirectional Support

Tex Texin

Internationalization Architect

Bidirectional (Bidi) Language Support

- HTML 4 DIR attribute

  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

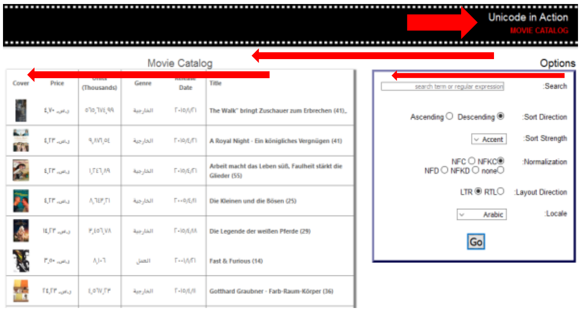
- HTML 5 – Isolates

  <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



Cover	Price	Units (Thousands)	Genre	Release Date	Title
	٤٫٧٠ ر.س.	[illegible]	[illegible]	٢٠١٥/١/٢١	The Walk" bringt Zuschauer zum Erbrechen (41)_
	٤٫٢٣ ر.س.	[illegible]	[illegible]	٢٠١٥/٤/٢١	A Royal Night - Ein königliches Vergnügen (41)
	٤٫٢٣ ر.س.	[illegible]	[illegible]	٢٠١٥/٤/٢١	Arbeit macht das Leben süß, Faulheit stärkt die Glieder (55)
	٤٫٢٣ ر.س.	[illegible]	[illegible]	٢٠٠٩/٤/١١	Die Kleinen und die Bösen (25)
	١٤٫٢٣ ر.س.	[illegible]	[illegible]	٢٠١٥/٤/١٨	Die Legende der weißen Pferde (29)
	٣٫٥٠ ر.س.	[illegible]	[illegible]	٢٠٠١/١/٢١	Fast & Furious (14)
	٢٤٫٢٣ ر.س.	[illegible]	[illegible]	٢٠١٥/٤/١١	Gotthard Graubner - Farb-Raum-Körper (36)

:Search search term or regular expression
:Sort Direction Ascending ○ Descending ●
:Sort Strength Accent
:Normalization NFC ○ NFKC ● NFD ○ NFKD ○ none ○
:Layout Direction LTR ● RTL ○
:Locale Arabic

Bidi References

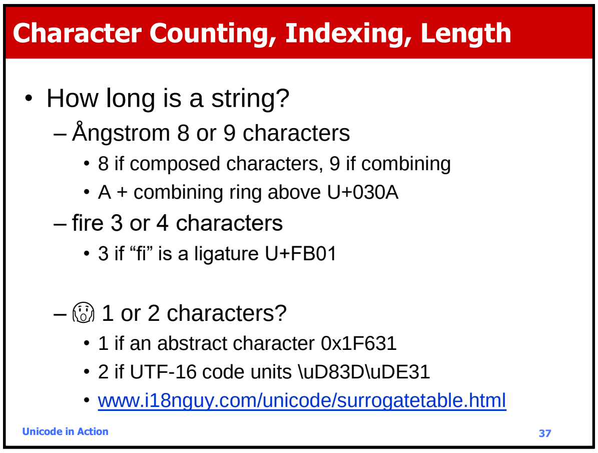
- W3C Bidi Tutorial
  - www.w3.org/International/tutorials/bidi-xhtml/
- Inline markup and bidirectional text in HTML
  - www.w3.org/International/articles/inline-bidi-markup/
- Additional Requirements for Bidi in HTML and CSS
  - www.w3.org/TR/html-bidi/
- Unicode Bidirectional Algorithm
  - (Unicode 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

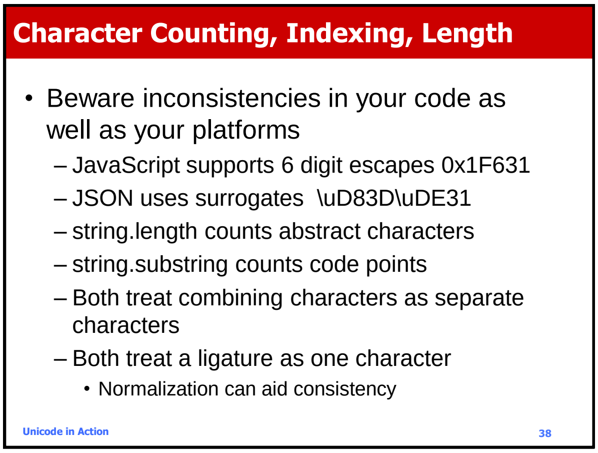
Character Counting, Indexing, Length

- 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

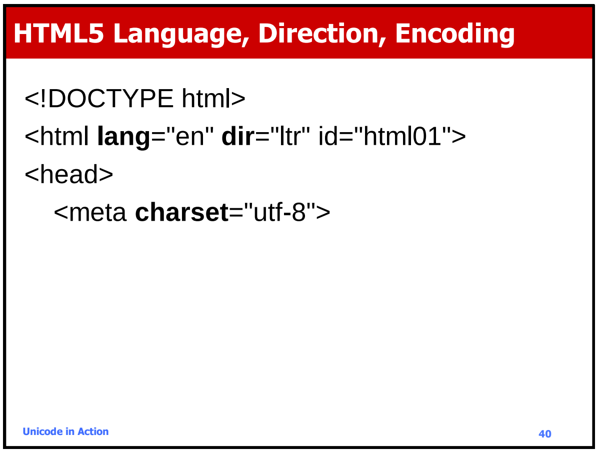


Character Counting, Indexing, Length

- 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

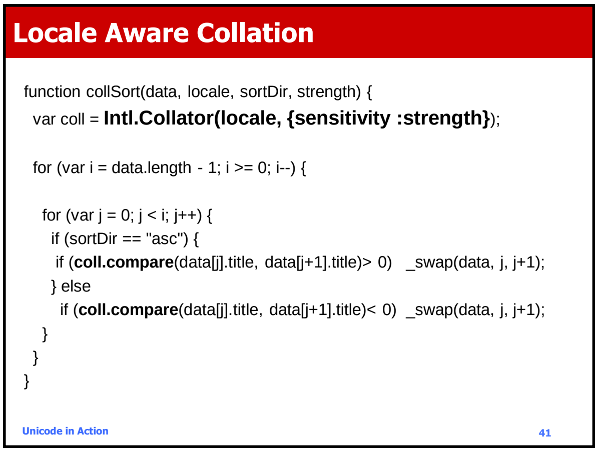
Internationalized Code

Tex Texin
Internationalization Architect



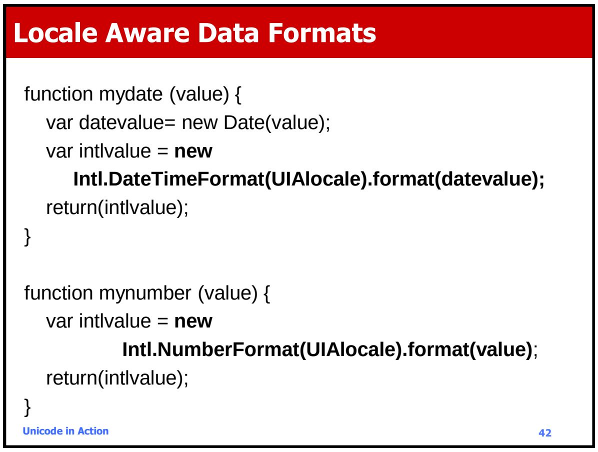
HTML5 Language, Direction, Encoding

```
<!DOCTYPE html>
<html lang="en" dir="ltr" id="html01">
<head>
    <meta charset="utf-8">
```



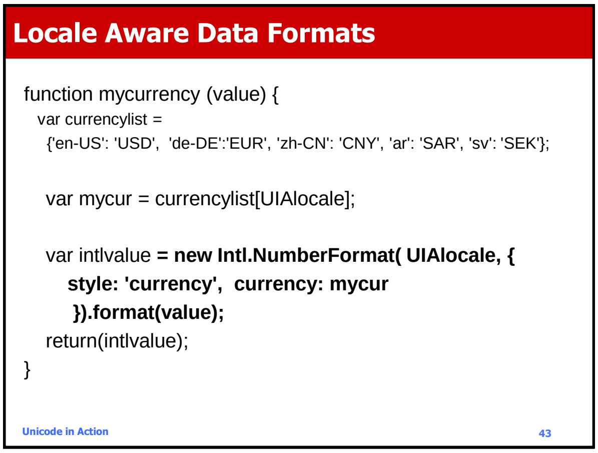
Locale Aware Collation

```
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);
    }
  }
}
```



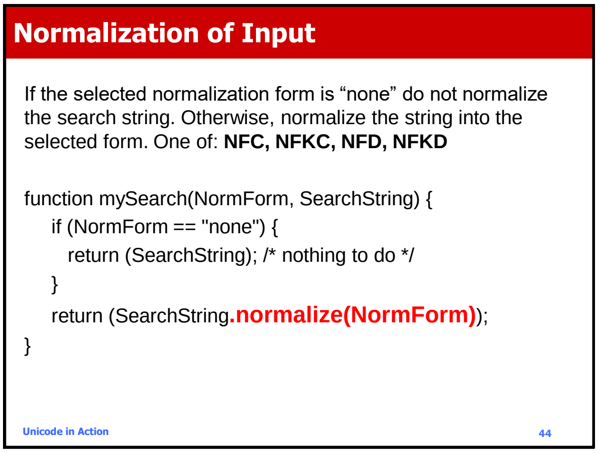
Locale Aware Data Formats

```
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);
}
```



Locale Aware Data Formats

```
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);
}
```

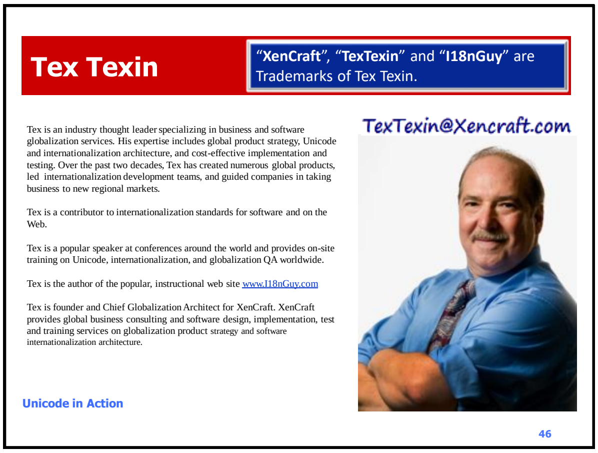


Normalization of Input

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**

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

Questions



Tex Texin

“**XenCraft**”, “**TexTexin**” and “**I18nGuy**” are Trademarks of Tex Texin.

TexTexin@Xencraft.com

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.