Oxford Internet Institute
University of Oxford

UNIVERSITY OF OXFORD

Country/Language Selectors Matter:

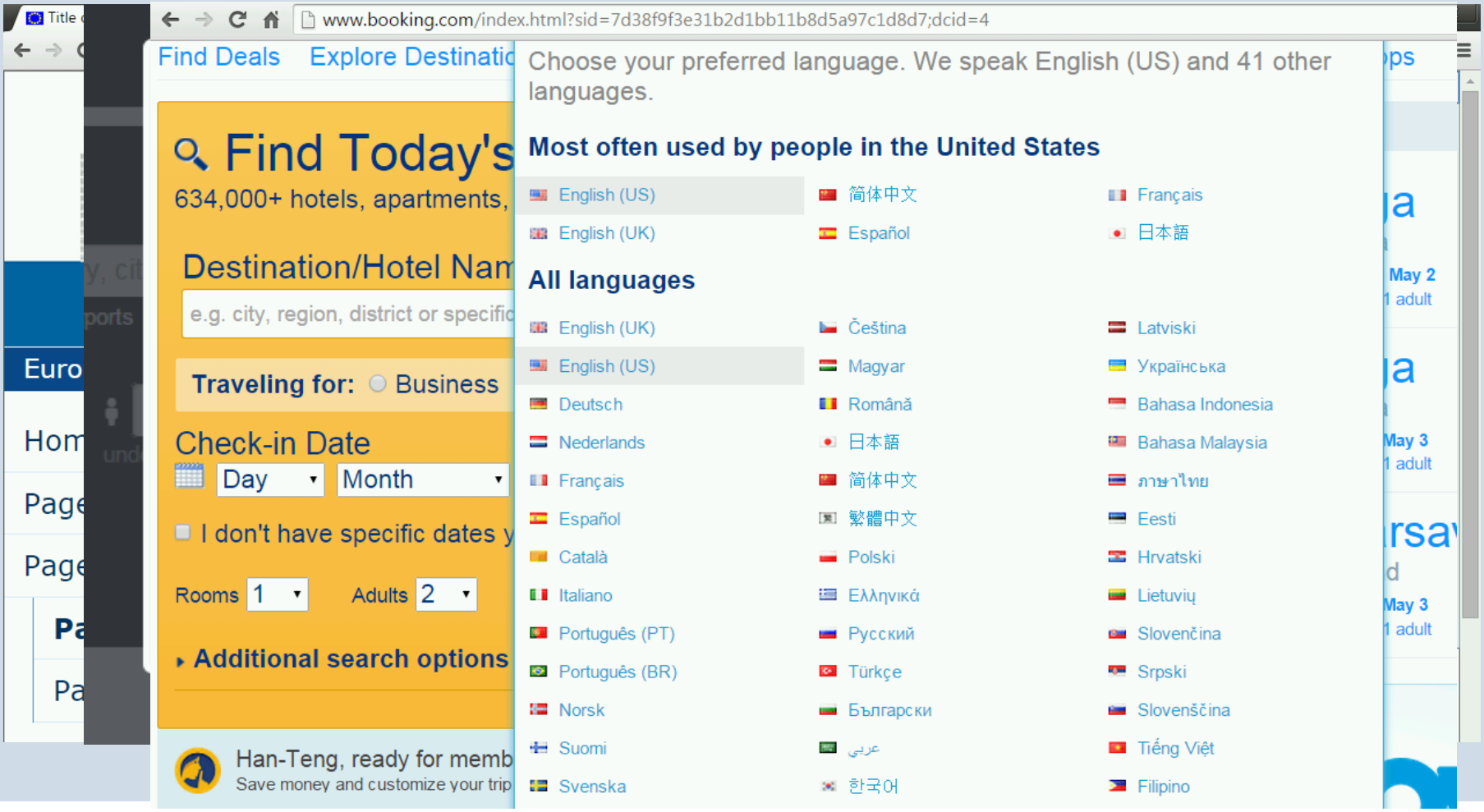
A Call to Build and Implement Common Translation and Country-Language Selector Repository

Han-Teng Liao
Oxford Internet Institute
University of Oxford

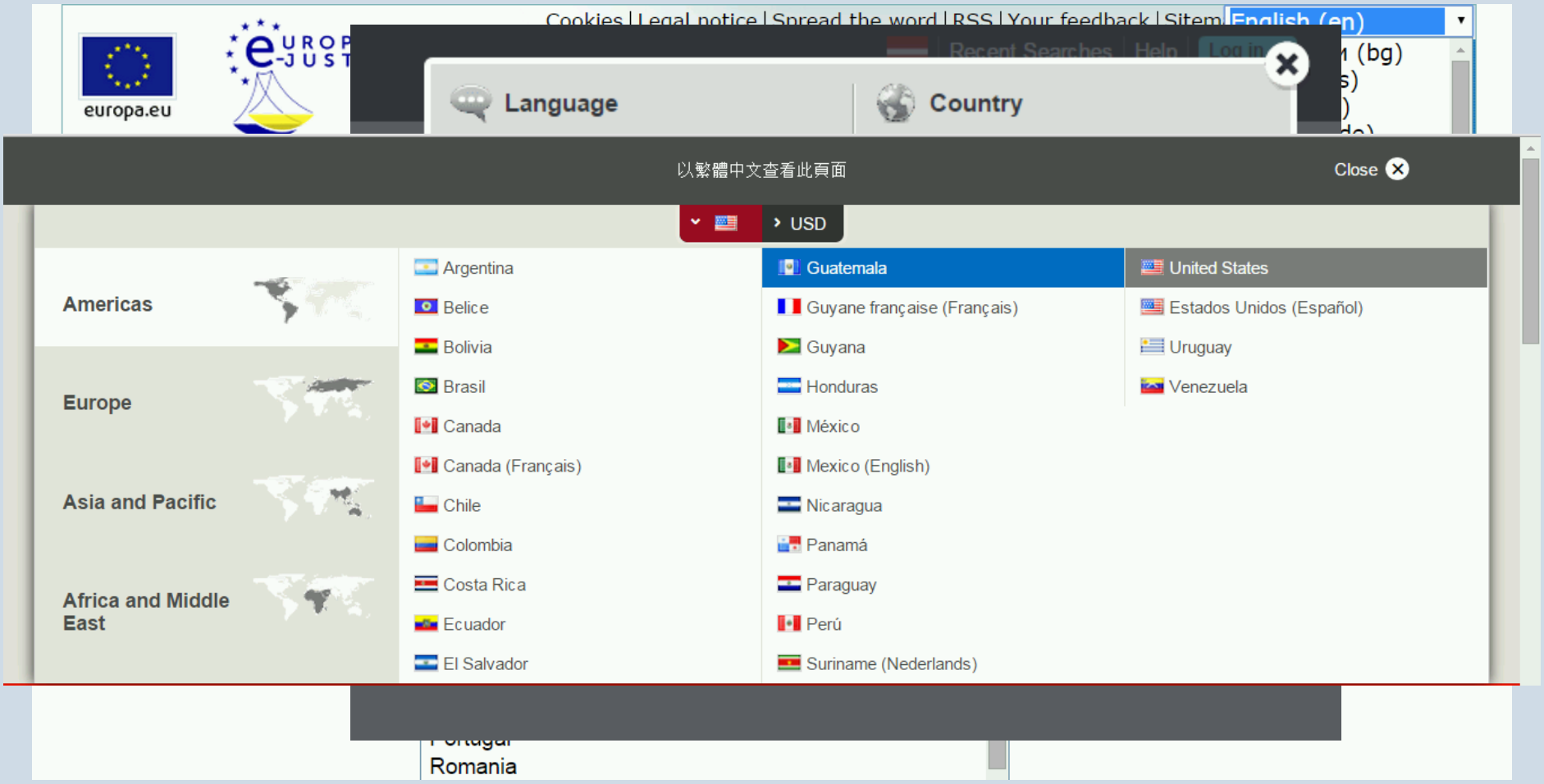
hanteng@gmail.com

For Riga Summit 2015

What is a Language Selector(or Switcher)?



What is a Country Selector (or Switcher)?



Why they matter?

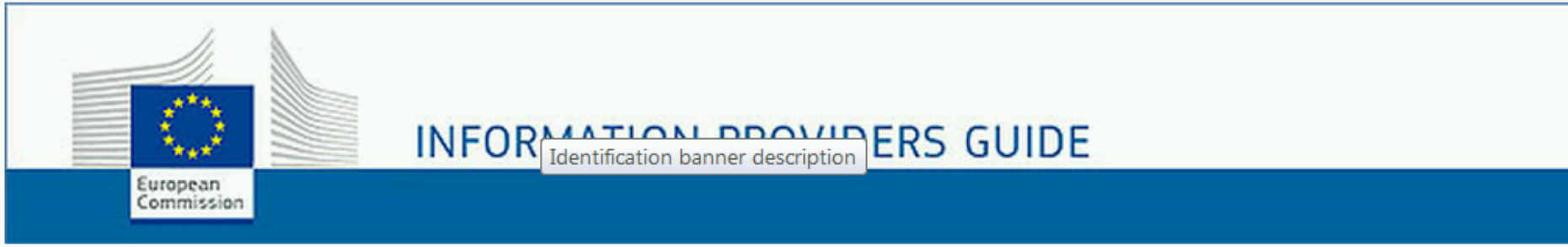
- First step
  - for users to select the preferred interface
- A paradox of choice
  - More options could mean higher cognitive cost for users to find the language they want
- European Commission's **Information Providers Guide** (IPG) states that

  "Language selection tool is **mandatory…**"

5. Language selection tool

Language selection tool is **mandatory** and must be present on all pages (even if the page exists in only one language). The language selection tool provides the only means of horizontal navigation between languages. It also provides an indication of which language versions of the page exist.

6. Identification banner



Why they matter?

…even within Chinese Wikipedia

字詞轉換 ×

本文使用公共轉換組「外國地名翻譯」。 [編輯][展開]

本文使用公共轉換組「英國政治人物」。 [編輯][展開]

本文使用全文手工轉換 [編輯]

- 简体：马斯特里赫特；繁體：馬斯特里赫特；台灣：馬斯垂克；當前顯示為：馬斯垂克
- 大陆：欧洲空间局；台灣：歐洲太空總署；當前顯示為：歐洲太空總署
- 简体：卡梅伦；繁體：卡梅倫；台灣：卡麥隆；香港：卡梅倫；澳門：卡梅倫；當前顯示為：卡麥隆
- 大陆：戴维·卡梅伦；台灣：大衛·卡麥隆；香港：大衛·卡梅倫；當前顯示為：大衛·卡麥隆
- 简体：戈登·布朗；大陆：戈登·布朗；台灣：戈登·布朗；香港：白高敦；新加坡：戈登·布朗；當前顯示為：戈登·布朗
- 简体：欧洲聯盟；大陆：欧洲联盟；台灣：歐洲聯盟；香港：歐盟；新加坡：欧洲联盟；澳門：歐盟；當前顯示為：歐洲聯盟

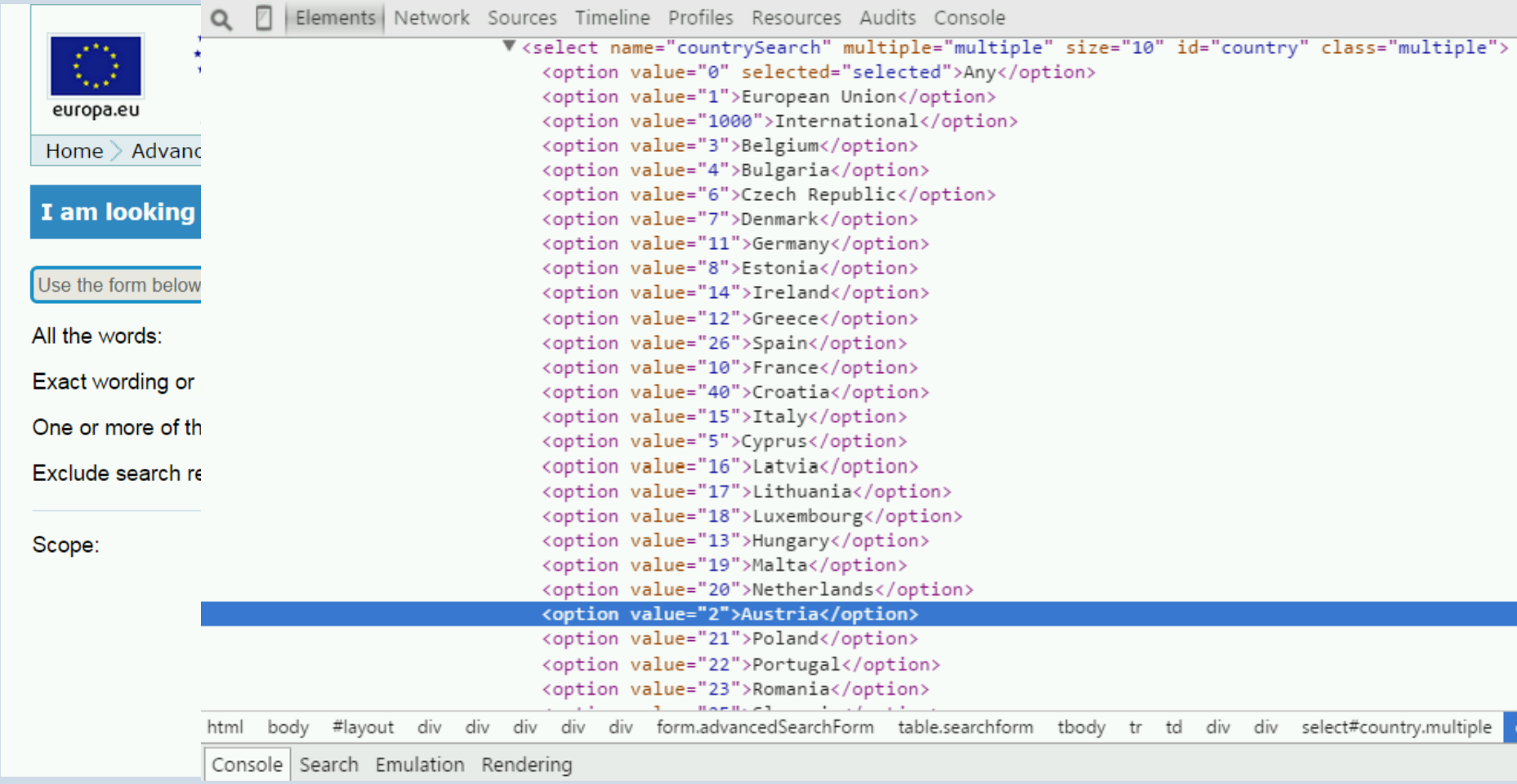
字詞轉換說明

字詞轉換是中文維基的一項自動轉換，目的是通過電腦程式自動消除繁簡、地區詞等不同**用字模式**的差異，以達到閱讀方便。字詞轉換包括全局轉換和手動轉換，本說明所使用的標題轉換和全文轉換技術，都屬於手動轉換。如果您想對我們的字詞轉換系統提出一些改進建議，或者提交應用面更廣的轉換（中文維基百科全站乃至MediaWiki軟體），或者報告轉換系統的錯誤，請前往Wikipedia:字詞轉換請求或候選發表您的意見。

PhD Data Intensive Work comparing Chinese Wikipedia and Baidu Baike

- All pages (2,500,000+) and all external links (2,000,000+) inside both encyclopedias
- 270,000 web links parsed and analysed based on 3000 queries of search engine result pages across nine Chinese search engine variants
- 60,000+ Sina Weibo and Twitter microblog posts
- Chinese and East Asian Internet Penetration Rates (historical data)
- Geographic distribution of power users in both encyclopedias

What’s wrong?


europa.eu
Home
I am looking
Use the form below
All the words:
Exact wording or
One or more of th
Exclude search re
Scope:
Elements Network Sources Timeline Profiles Resources Audits Console
<select name="countrySearch" multiple="multiple" size="10" id="country" class="multiple">
<option value="0" selected="selected">Any</option>
<option value="1">European Union</option>
<option value="1000">International</option>
<option value="3">Belgium</option>
<option value="4">Bulgaria</option>
<option value="6">Czech Republic</option>
<option value="7">Denmark</option>
<option value="11">Germany</option>
<option value="8">Estonia</option>
<option value="14">Ireland</option>
<option value="12">Greece</option>
<option value="26">Spain</option>
<option value="10">France</option>
<option value="40">Croatia</option>
<option value="15">Italy</option>
<option value="5">Cyprus</option>
<option value="16">Latvia</option>
<option value="17">Lithuania</option>
<option value="18">Luxembourg</option>
<option value="13">Hungary</option>
<option value="19">Malta</option>
<option value="20">Netherlands</option>
<option value="2">Austria</option>
<option value="21">Poland</option>
<option value="22">Portugal</option>
<option value="23">Romania</option>
html body #layout div div div div div form.advancedSearchForm table.searchform tbody tr td div div select#country.multiple
Console Search Emulation Rendering

Inconsistent choice between names

- Official names
  - "Korea, Republic of“ vs “Republic of Korea”
  - “China, Republic of” vs “Republic of China”
  - “Taiwan, Province of China”
- Customary names
  - South Korea
  - Taiwan
  - Taiwan
- What’s in a name? Country names are technical, cultural and political.

Collations

Collation

From Wikipedia, the free encyc

This article is about colla
Collation (disambiguatior

Collation is the assembly of based on numerical order or fundamental element of mos

Collation differs from *classifi* logical categories, while colla on the form of their identifier a set of possible identifiers, of items of information (items

A collation algorithm such as comparing two given charact order has been defined in thi items into that order.

10.1.2 Character Sets and Collations in MySQL

The MySQL server can support multiple character sets. To list the available character sets, use the `SHOW CHARACTER SET` statement. A partial listing follows. For more complete information, see Section 10.1.13, "Character Sets and Collations That MySQL Supports".

```
mysql> SHOW CHARACTER SET;
+----------+-----------------------------+---------------------+--------+
| Charset  | Description                 | Default collation   | Maxlen |
+----------+-----------------------------+---------------------+--------+
| big5     | Big5 Traditional Chinese    | big5_chinese_ci     |      2 |
| dec8     | DEC West European           | dec8_swedish_ci     |      1 |
| cp850    | DOS West European           | cp850_general_ci    |      1 |
| hp8      | HP West European            | hp8_english_ci      |      1 |
| koi8r    | KOI8-R Relcom Russian       | koi8r_general_ci    |      1 |
| latin1   | cp1252 West European        | latin1_swedish_ci   |      1 |
| latin2   | ISO 8859-2 Central European | latin2_general_ci   |      1 |
| swe7     | 7bit Swedish                | swe7_swedish_ci     |      1 |
| ascii    | US ASCII                    | ascii_general_ci    |      1 |
| ujis     | EUC-JP Japanese             | ujis_japanese_ci    |      3 |
| sjis     | Shift-JIS Japanese          | sjis_japanese_ci    |      2 |
| hebrew   | ISO 8859-8 Hebrew           | hebrew_general_ci   |      1 |
| tis620   | TIS620 Thai                 | tis620_thai_ci      |      1 |
| euckr    | EUC-KR Korean               | euckr_korean_ci     |      2 |
| koi8u    | KOI8-U Ukrainian            | koi8u_general_ci    |      1 |
| gb2312   | GB2312 Simplified Chinese   | gb2312_chinese_ci   |      2 |
| greek    | ISO 8859-7 Greek            | greek_general_ci    |      1 |
| cp1250   | Windows Central European    | cp1250_general_ci   |      1 |
| gbk      | GBK Simplified Chinese      | gbk_chinese_ci      |      2 |
| latin5   | ISO 8859-9 Turkish          | latin5_turkish_ci   |      1 |
...
```

Any given character set always has at least one collation. It may have several collations. To list the collations for a

What are the possible solutions?

- Unicode Collation Charts:
  - http://www.unicode.org/charts/uca/
- Unicode CLDR
  - Preferences on Customary names over Official names

Suggestions

- Consistent collation in Web design
- Autocomplete based on both codes and names
- Complete world coverage for templates
  - EU official language and member state subsets
  - Geographic categorization schemes
- Separation of languages and states (flags)

Source: http://flagsarenotlanguages.com/blog/

Best practice for prese ×

flagsarenotlanguages.com/blog/best-practice-for-presenting-languages/

Flags are not languages

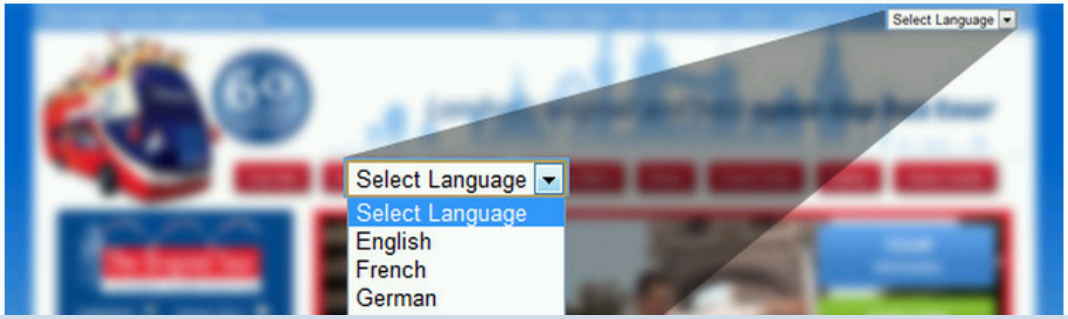
A blog about designing global user experiences: beyond language, location & culture.

Best practice for presenting languages

When presenting links to content in different languages, consider the following:

Always use the name of the language in its local format

If you're linking to a page in German, label it as 'Deutsch' — not 'German'.



Resources

- Why flags do not represent languages
- Best practice for presenting languages
- Iconography for translations: best practice for communicating availability of translated content
- About this site

Recent Posts

- Post Office Corporate and ignoring Scotland
- Babbel.com: regional language variations and flags
- memegenerator.net and overcomplicating language input
- Case study: onefinestay.com and

Source: http://flagsarenotlanguages.com/blog/

Implications

- Best practices needed
  - for presenting a list of languages (and countries)
  - for users to select their target language (and country)

- Pooling resources

- Information literacies promoted
  - on the use of language codes and country codes
  - on the difference between *customary* vs. *official* names

Oxford Internet Institute
University of Oxford

UNIVERSITY OF OXFORD

Thank you

PhD Data Intensive Work comparing Chinese Wikipedia and Baidu Baike

- All pages (2,500,000+) and all external links (2,000,000+) inside both encyclopedias
- 270,000 web links parsed and analysed based on 3000 queries of search engine result pages across nine Chinese search engine variants
- 60,000+ Sina Weibo and Twitter microblog posts
- Chinese and East Asian Internet Penetration Rates (historical data)
- Geographic distribution of power users in both encyclopedias

Percentage of visibility scores: encyclopedia sites among the top-10