Internationalization & Localization on the Web

Olle Olsson
Swedish W3C Office
Swedish Institute of Computer Science (SICS)

"Språk och Internet" – Stockholm, Sweden, May 14, 2008

SICS – Swedish Institute of Computer Science

National research institute
- R&D in information and communication technologies

Objective:
- conduct advanced and focused research in strategic areas of computer science

Sponsors:
TeliaSonera, Ericsson,
Saab Systems,
FMV (Defence Materiel Administration),
Green Cargo (Swedish freight railway operator),
ABB,
Bombardier Transportation

The Internet vs. the Web

- The Internet
  - Basic technical infrastructure – plumbing
  - Shuffling bits
  - Content agnostic
- The Web
  - Set of technologies
  - Processing content
  - Content-dependent

Languages and the Web

- A global web
- Many (natural) languages at *international* scale
- Global reachability
- Taking advantage ...

- Many (natural) languages at *national* scale
- Society-for-all

- Heterogeneity (language, culture, ...)
  - Problem?
  - Opportunity?

Who cares?

- Purchase power
  - Users are three (3) times more likely to buy a product when they are addressed in their own language
    - Source: “Strategies for Global Sites” Donald DePalma, Forrester Research, Inc.
- Customer service
  - Customer service costs drop when instructions are displayed in a user’s native language
    - Source: “Strategies for Global Sites” Donald DePalma, Forrester Research, Inc.
- Increased revenue
  - One large IT company discovered that a significant percentage of inquiries were coming from South Korea - they created a Korean website and revenues rose by 8 percent
    - Source: “Global eCommerce” Donald J. Plumley, Bowne Global Solutions

Challenges

- Technological support for languages of the world
  - Representation
  - Presentation
- Support for multi-lingual content
- Predicting rendering on devices
- Effective production of contents
- Effective adaptation of contents

Web as a platform for multi-*

- Set of web standards for content and processing
- Coherent set of standards
- Standards adopted by users
- Standards supported by vendors

- World Wide Web Consortium (W3C)
  - Consensus-based standards development
  - Work targets critical areas of needs

- W3C Internationalization work

Internationalization vs. Localization

- Internationalization (I18N)
  - The design and development of a product, application or document content that enables easy *localization* for target audiences that vary in culture, region, or language.

- Localization (L10N)
  - The adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific target market

What about I18N?

- Provides technology for features that facilitate local and international access:
  - bi-directional (bidi) text
  - language identification
  - vertical text
  - non-Latin typography
  - etc
- Provides technology for features relating to local, regional, language or culturally related concerns:
  - date/time formats
  - calendar localization
  - number formats & numeral systems
  - personal names & forms of address
  - etc

What about L10N?

- Site customization related to:
  - Translation of languages
    - Manual; automatic; ....
  - Numeric, date and time formats
  - Use of currency
  - Keyboard usage
  - Symbols, icons and color
  - Sensitivity to cultural perceptions in regards to language and visual images
  - etc.

I18N - Technological needs

Need to identify content characteristics:

... natural language

... character set

... encoding

Need to identify delivery context characteristics:

... user preferences

... device properties

Characters – character sets and encodings

European alphabetic scripts
Latin
Greek
Cyrillic
Armenian
Georgian
Runic
Ogham
Modifier letters
Combining characters

East Asian scripts
Han
Hiragana
Katakana
Hangul
Bopomofo
Yi

Middle East scripts
Hebrew
Arabic
Syriac
Thaana

Symbols
Currency symbols
Letter like symbols
Mathematic operators
Numeric forms
Technical symbols
Geometrical symbols
Miscellaneous symbols & dingbats
Enclosed & square
Braille

South & South East Asian scripts
Devanagari
Bengali
Gurmukhi
Gujurati
Panjabi
Oriya
Tamil
Telugu
Kannada
Malayalam
Sinhala
Thai
Lao
Tibetan
Myanmar
Khmer

Additional scripts
Ethiopic
Cherokee
Canadian Aboriginal Syllabics
Mongolian
Tifinagh

Etc....

Characters – character sets and encodings

		A	א	好	不
Code point		41	5D0	597D	233B4
Encodings	UTF-8	41	D7 90	E5 A5 BD	F0 A3 8E B4
	UTF-16	00 41	05 D0	59 7D	D8 4C DF B4
	UTF-32	00 00 00 41	00 00 05 D0	00 00 59 7D	00 02 33 B4

Characters

- Declaration of encoding
  - `Content-Type: text/html; charset=iso-8859-1`

雪 zh-Hans

雪 zh-Hant

雪 ja

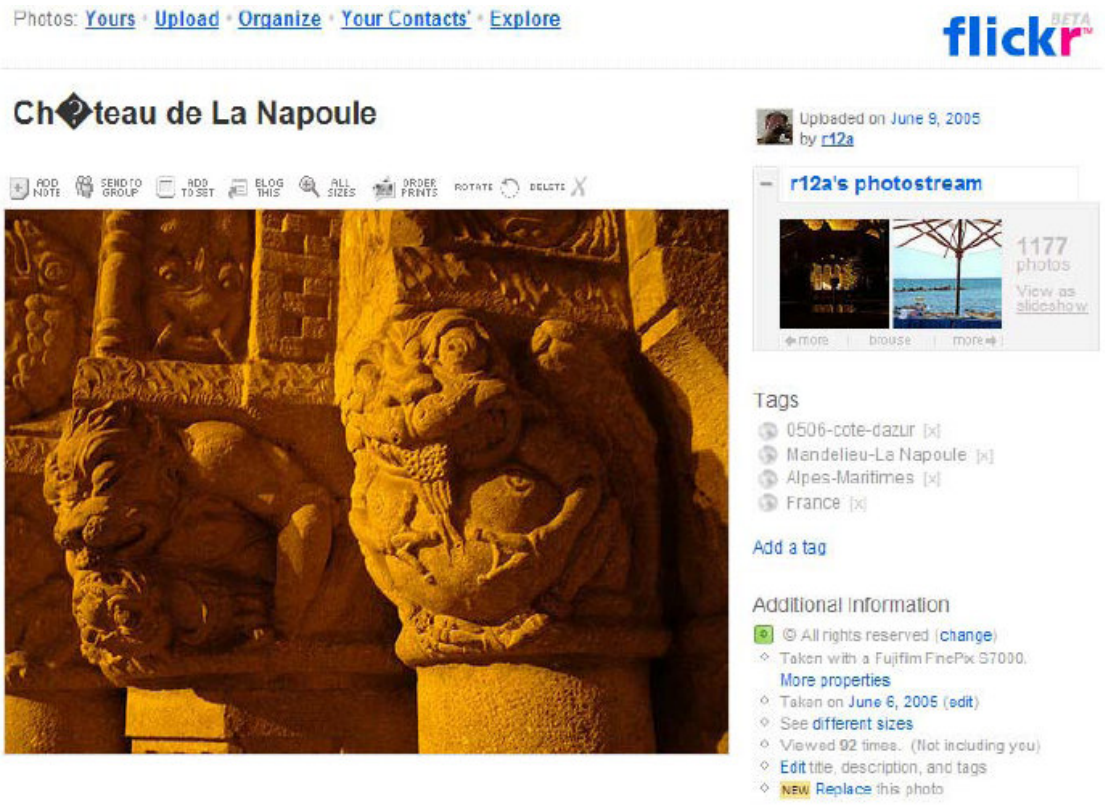
雪 ko

- Ruby annotations

しんかんせん ← *ruby text*

新幹線 ← *ruby base*

Characters – character sets and encodings



IRIs: Internationalized URIs

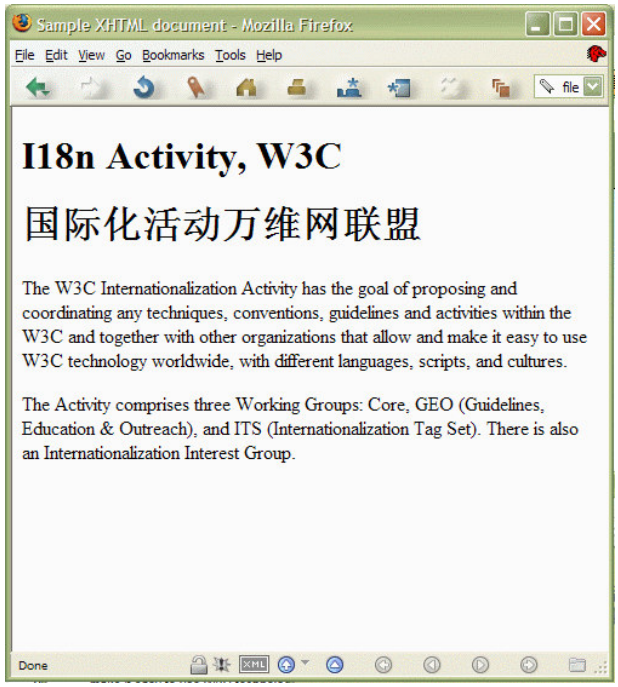
- Internationalized Domain Names in IRIs
  - IRI: http://räksmörgås.josefsson.org,
    http:// 日本語 .w3.mag.keio.ac.jp
  - URI (forward-looking):
    http://r%C3%A4ksm%C3%B6rg%C3%A5s.josefsson.org,
    http://%E7%B4%8D%E8%B1%86.w3.mag.keio.ac.jp
  - URI (backwards-compatible):
    http://xn--rksmrgs-5wao1o.josefsson.org,
    http://xn--99zt52a.w3.mag.keio.ac.jp

Separation content & presentation

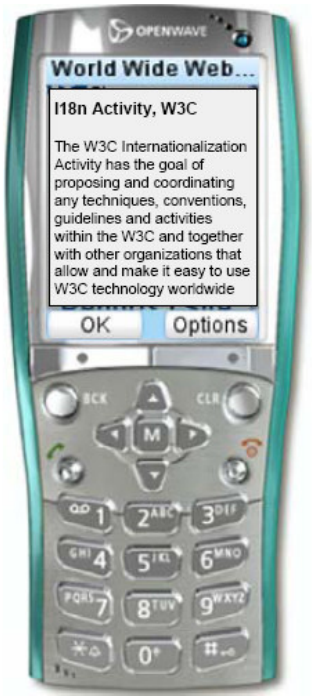
```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
   "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
   <head>
      <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
      <title>About the W3C I18n Activity</title>
      <style type="text/css" src="mystyling.css" />
      </head>

    <body>
      <h1>I18n Activity, W3C</h1>
      <div class="international-text" xml:lang="zh-Hans"
                  lang="zh-Hans">国际化活动万维网联盟</h1>
      <div class="description">
         <p>The W3C Internationalization Activity has the goal of proposing
            and coordinating any techniques, conventions, guidelines and
            activities within the W3C and together with other organizations
            that allow and make it easy to use W3C technology worldwide,
            with different languages, scripts, and cultures.</p>
         <p>The Activity comprises three Working Groups: Core, GEO
            (Guidelines, Education & Outreach), and ITS (Internationalization
            Tag Set). There is also an Internationalization Interest Group.</p>
         </div>
       </body>
   </html>
```



Separation ... style sheets



Document formats: content language

HTTP Content-Language header

```
HTTP/1.1 200 OK
Date: Wed, 05 Nov 2003 10:46:04 GMT
Server: Apache/1.3.28 (Unix) PHP/4.2.3
...
Content-Type: text/html; charset=utf-8
Content-Language: en
```

Language attribute on html tag

Content-Language meta tag

Language attribute on embedded element

```
<html lang="en">
<head>
...
<meta http-equiv="Content-Language" content="en" />
...
</head>
<body>
<p>The French word for <em>cat</em> is
<em lang="fr">chat</em>.</p>
...
</body>
</html>
```


Document formats: speech synthesis

這一晚會如常舉行

這一|晚會|如常|舉行 — This banquet is held as usual.

這一|晚會|如|常|舉行 — If this banquet is held frequently.

這一晚|會|如常|舉行 — (An event) will be held tonight as usual.

Presentation: typography

当世界需要沟通时，请用Unicode。将于3月10日-12日在德国 Mainz 市举行的第十届统一码国际研讨会现在开始注册。本次会议将汇集各方面的专家。涉及的领域包括：国际互联网和统一码，国际化和本地化，统一码在操作系统和应用软件中的实现，字型，文本格式以及多文种计算等。

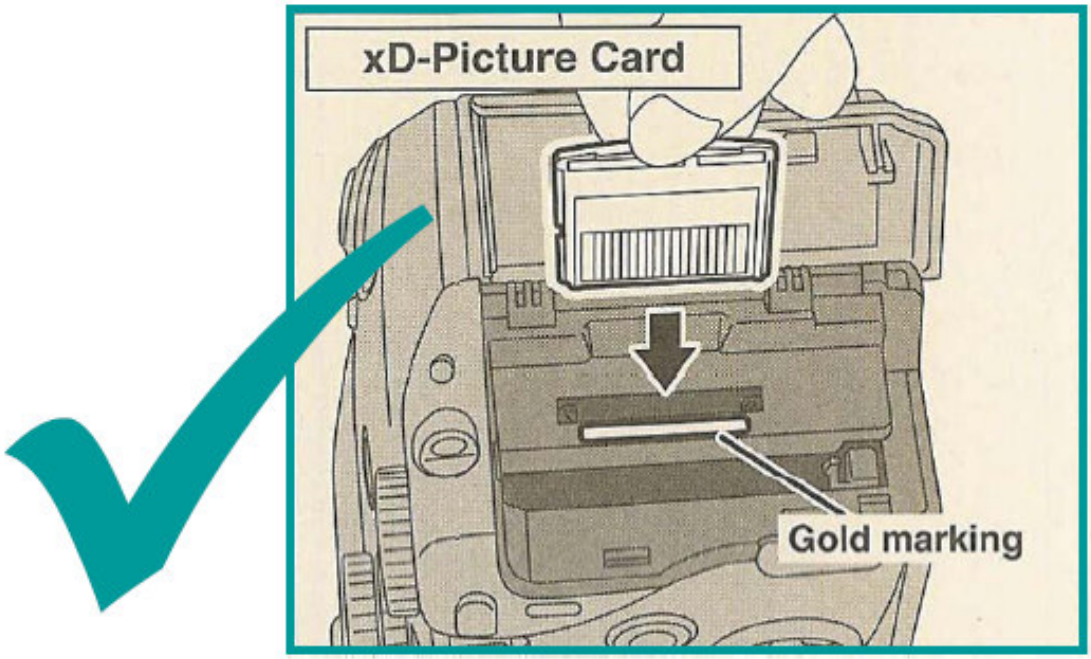
Presentation: layout

Cultural differences

Россия
г. Пермь 614055
ул. Крупской 93-82
Селивановой Юлии

Country:	United States
First name:	
Last name:	
Address:	
City:	
State:	AZ
Zip code:	
Telephone:	( )
Application date:	

Cultural differences



W3C Recommendations & informative docs

- Character Model for the World Wide Web 1.0: Fundamentals
- Internationalization Tag Set (ITS)
- Internationalized Resource Identifiers
- Ruby Annotation
- ...

- Internationalization Best Practices
- Specifying Language in XHTML & HTML Content
- Best Practices for XML Internationalization
- Unicode in XML and Other Markup languages
- ...

W3C I18N work

Internationalization Activity

- Core Working Group
  - Reviews, advice, and internationalization specifications
- ITS (Internationalization Tag Set) Working Group
  - Elements and attributes for schema developers
- GEO (Guidelines, Education & Outreach) Working Group
  - Making internationalization aspects of W3C technology better understood and more widely and consistently used

Summary

Internationalization means:

- using a Quality approach to reduce the overall cost and time to market/release of multinational deliverables
- designing into the product an internationalized base, and a modular and easily adaptable architecture
- not always doing extra work maybe just working in a better way