The multilingual web

Felix Sasaki

DFKI / University of Appl. Sciences Potsdam

W3C German-Austrian Office

felix.sasaki@dfki.de

Slides:

http://www.sasakiatcf.com/felix/publications/sasaki-webtechcon2010.pdf

About me

- Studied Japanese and Linguistics in Germany and Japan
- PhD in Computational Linguistics with a focus on Web technology & multilingual data
- 2005-2009: Work in Japan within the W3C Internationalization Activity
- Since 2009: Professor at Univ. of Appl. Sciences Potsdam / Manager of the W3C German-Austrian Office
- Since Autumn 2010: Senior Researcher at DFKI (German Research Center for Artificial Intelligence)

About W3C Offices

- A contact point for whose who
  - Don't know W3C very well (yet)
  - Want to ask specific questions like "Who is working on topic ABC …"
  - Are considering new web-related standardization work and are wondering where the best place to do that might be
- So for any questions related to the topics mentioned above, please don't care to bother me too ☺

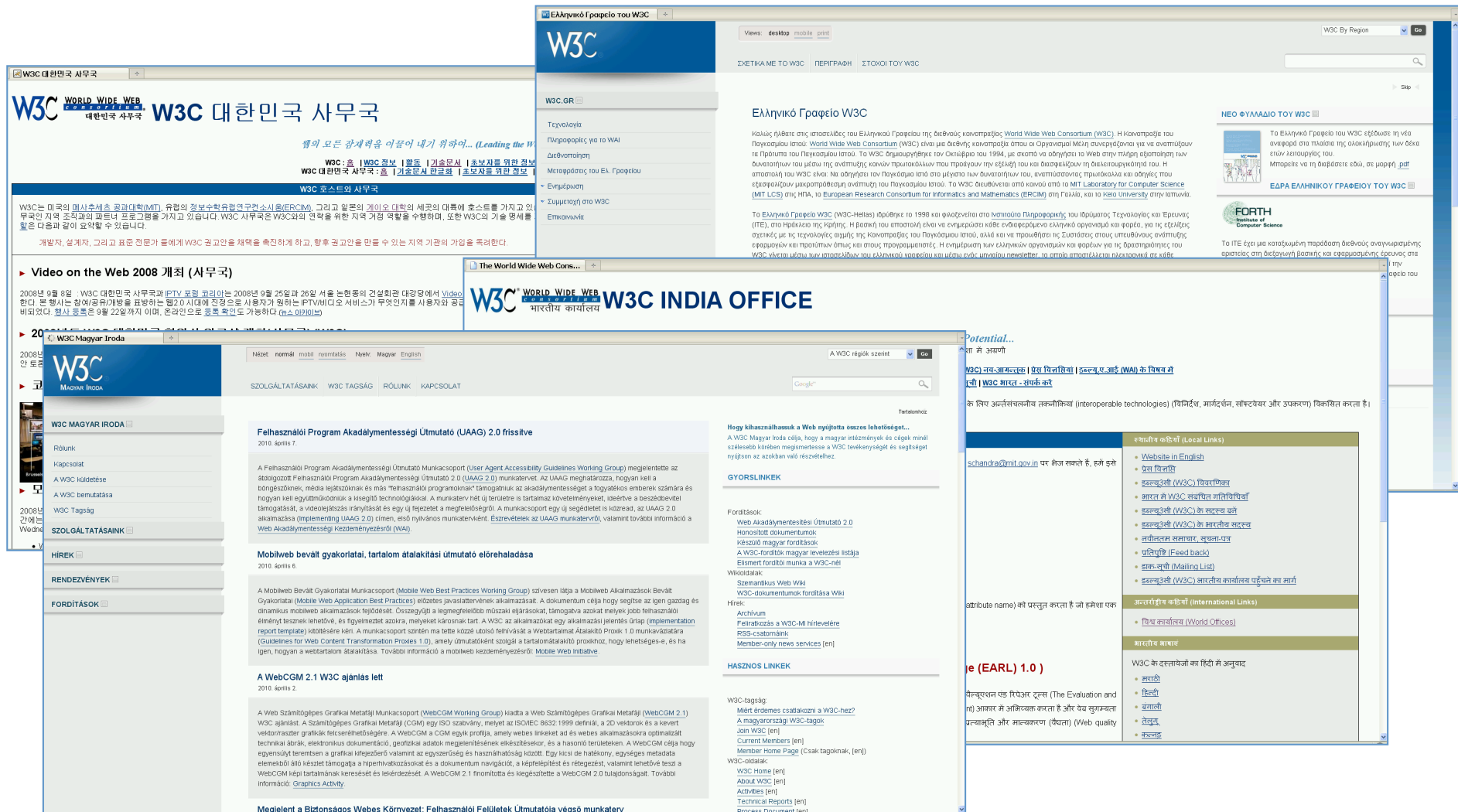# Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- Projects "Multilingual Web" and "META-NET"

Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- Projects "Multilingual Web" and "META-NET"

The Multilingual Web: <u>Content</u> in many languages and scripts

11. - 13. Oktober
Rheingoldhalle, Mainz

Localized Services

Devices everywhere, for everybody

Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- Projects "Multilingual Web" and "META-NET"

Localization (L10N): Give users what they need
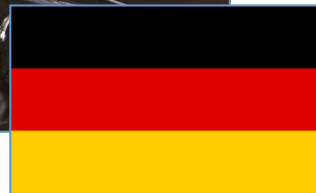
Internationalization (I18N):
Prepare yourself for Localization
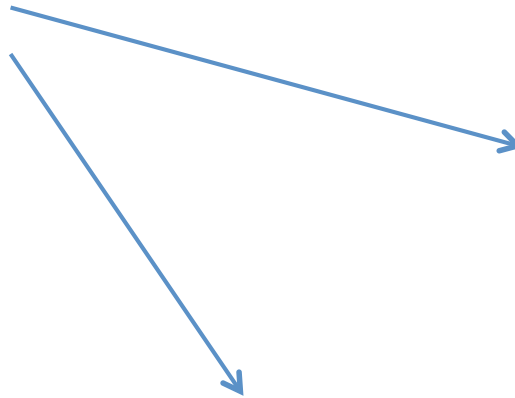
Proper I18N and L10N: Required for <u>successful</u> Globalization

Traditional topics of I18N on the Web

Traditional topics of I18N on the Web

- Use of Unicode in web technologies

Traditional topics of I18N on the Web

Internationalized Domain Names (IDN).

webtech
CONFERENCE 2010

11. - 13. Oktober
Rheingoldhalle, Mainz
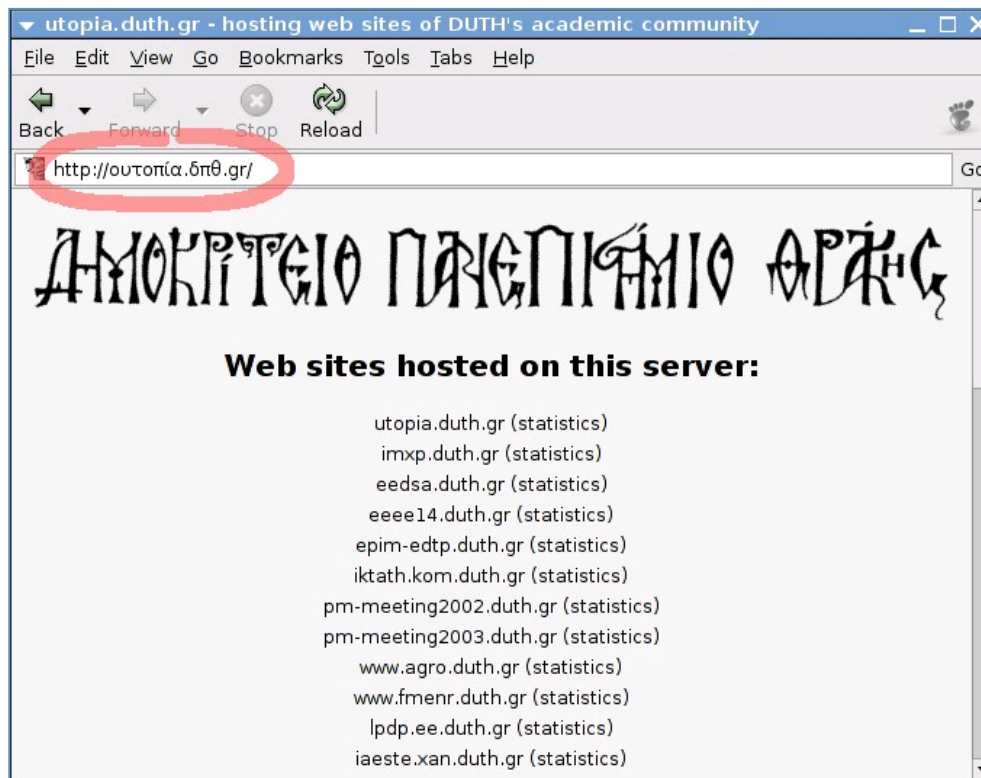
Traditional topics of I18N on the Web

Internationalized Domain Names (IDN). "Actually" possible since 2003, but <u>widely announced</u> last year

Traditional topics of I18N on the Web

- Internationalized Resource Identifier (IRI)
- I18N in <u>the path</u> of a Web Address, e.g.
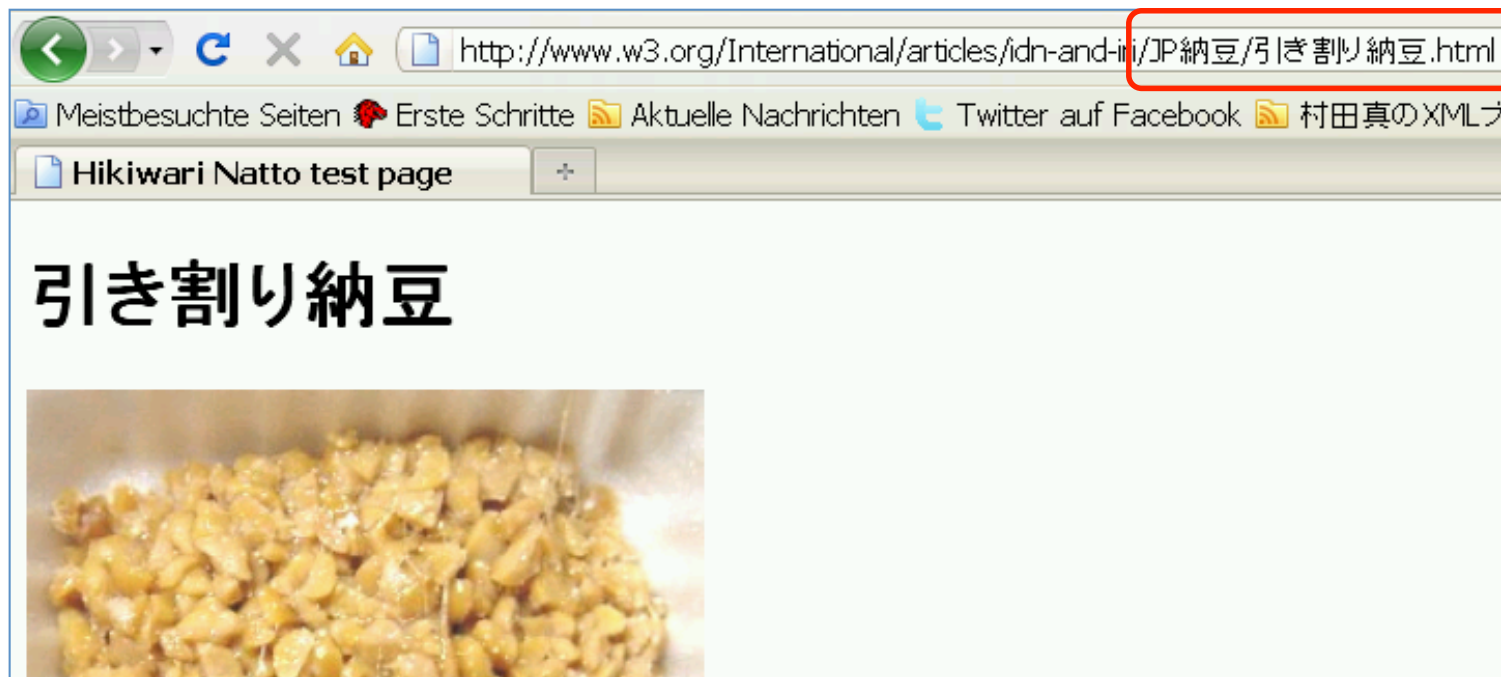
Traditional topics of I18N on the Web

- Language tags like "en", "en-us", "de", "ja", ...

Used e.g. for
content-negotation:
Give users what
(their browser says) they want
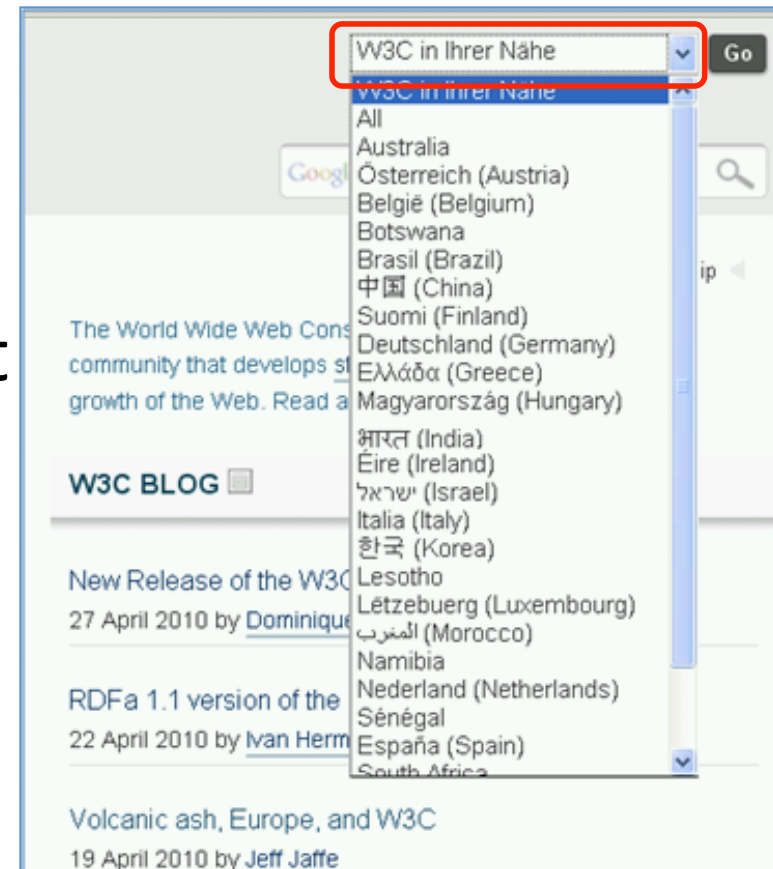
Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- Projects "Multilingual Web" and "META-NET"

What is the long tail?

- Making money by selling small, but <u>many</u> products
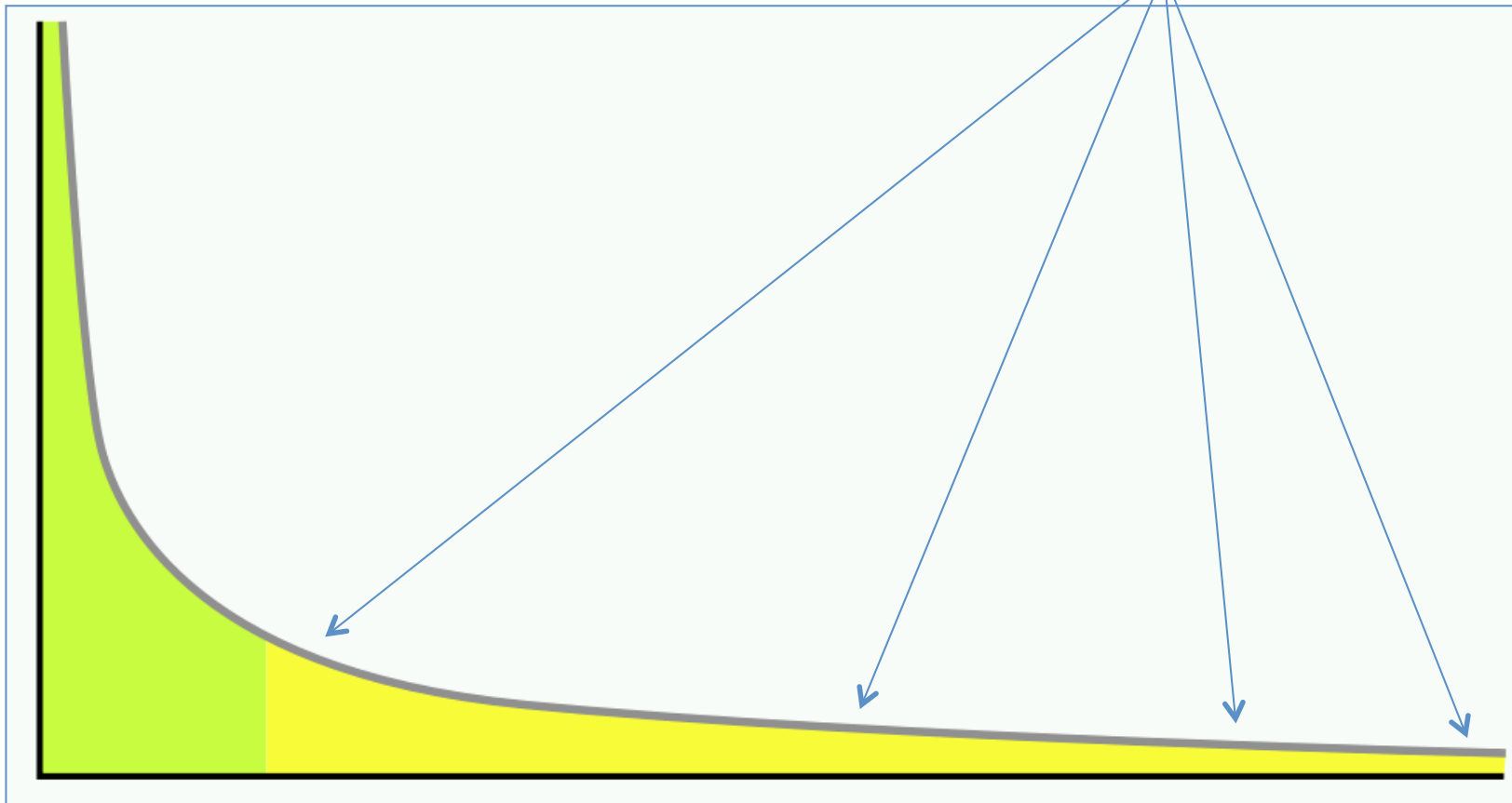
What is the long tail?

- Making money by selling small, but <u>many</u> products
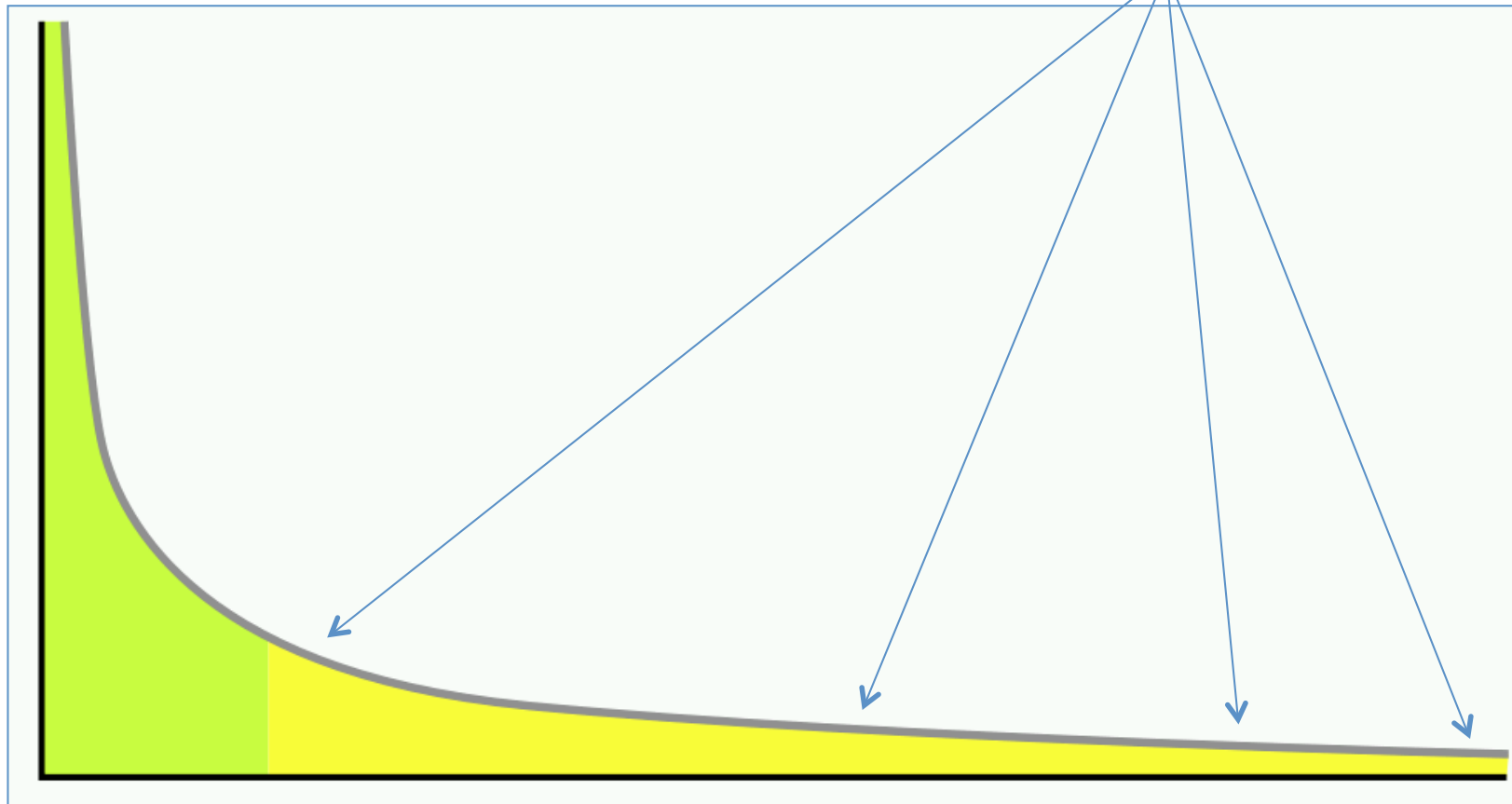
Example Amazon: "We sold more books today that didn't sell at all yesterday than we sold today of all the books that did sell yesterday."

What is the long tail?

- Making money by selling small, but <u>many</u> products
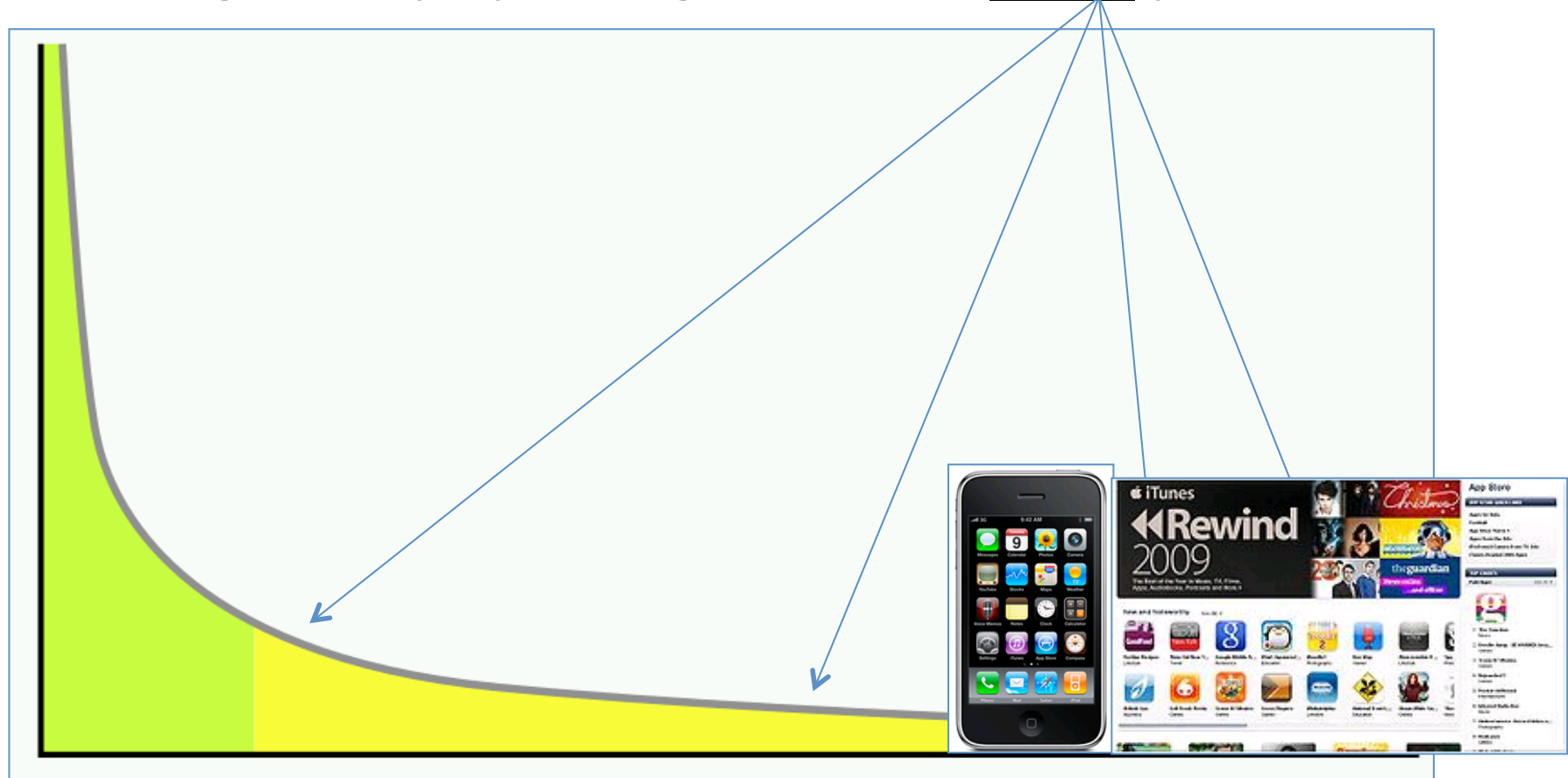
Example iPhone applications: Third-party applications like games, reference, GPS navigation, social networking, advertising for television shows, etc.

The long tail in I18N / l10N

- More specific content and services for <u>many , many devices and users</u>

The long tail in I18N / l10N

- More specific content and services for <u>many , many devices and users</u>
- Centralized I18N / L10N is too expensive for this scenario …
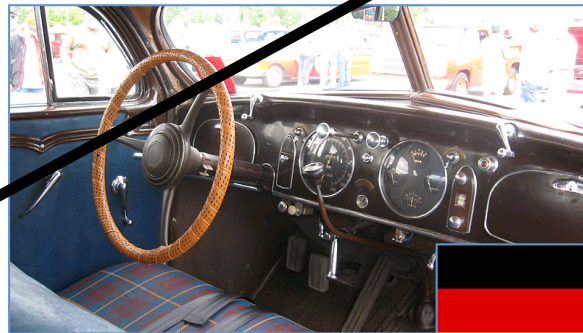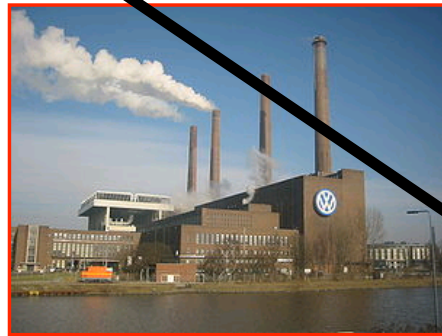
A path to a solution: have a look at how linked data works!
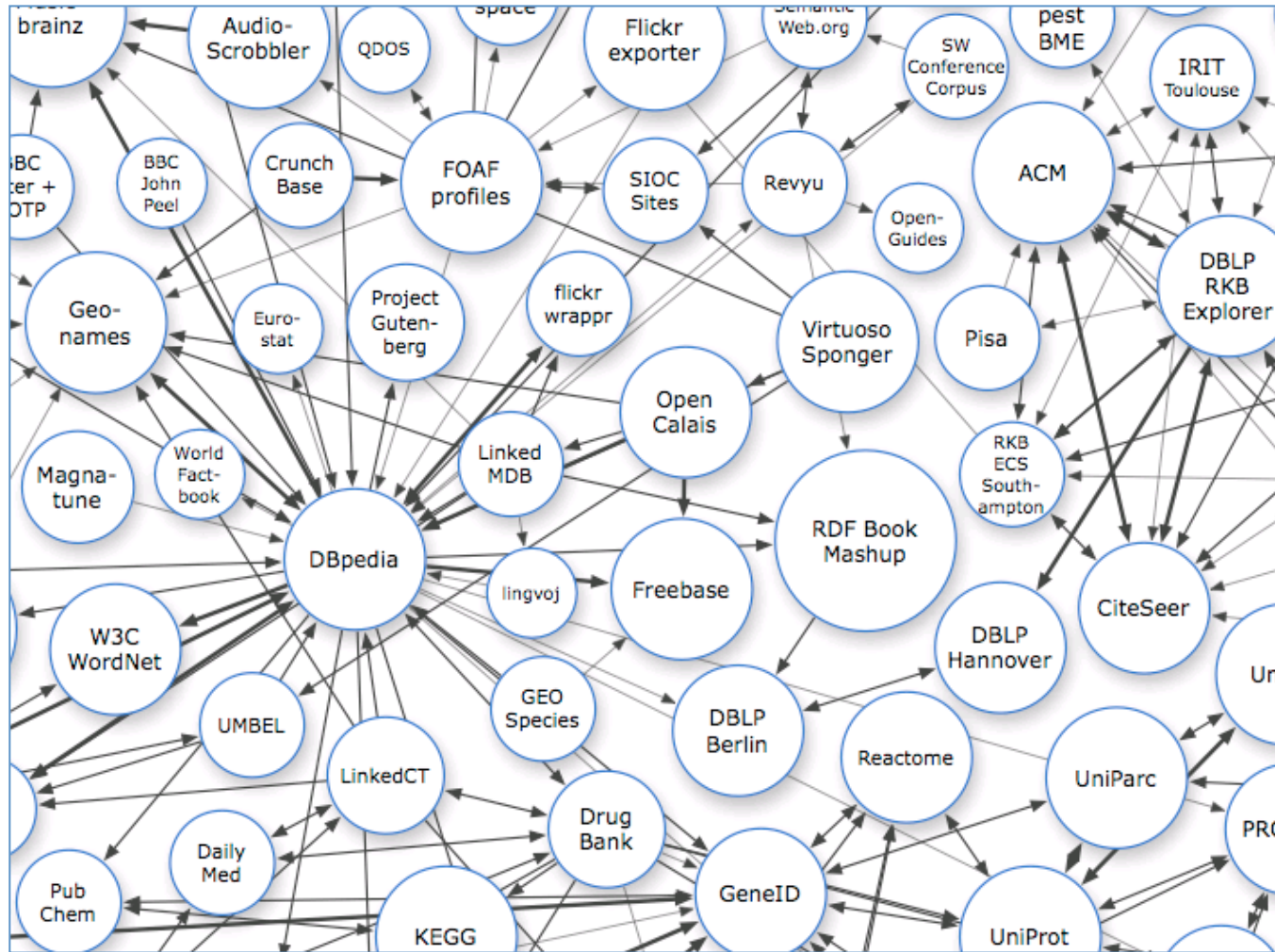
Reaching the long tail in I18N / I10N

- "Linked data" shows a path how to reach the "long tail" market
- Give people a means to create the synergies needed for the long tail, via enhanced, <u>standardized technologies</u>!
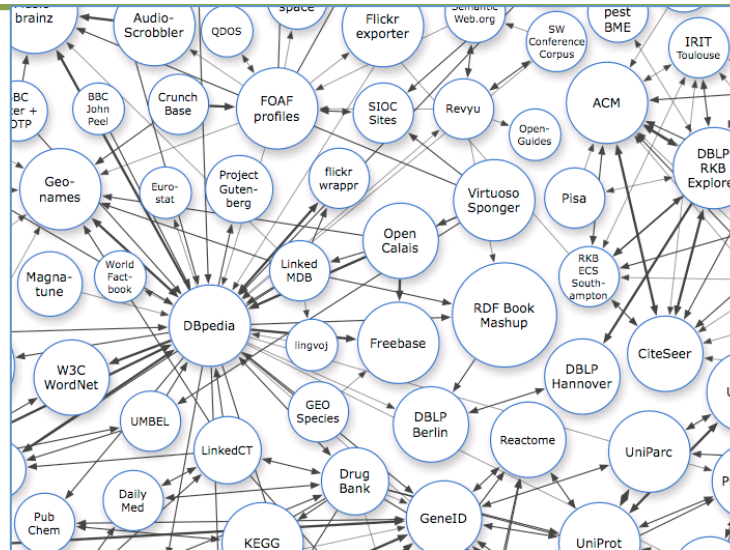
Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- The project "Multilingual Web"

Relatively new possibilities of language tags: what does this mean?

- de-1901
- de-1996
- ja-latn
- ja-latn-hepburn
- ja-latn-hepburn-heploc
- ase
- sgn

Relatively new possibilities of language tags: what does this mean?

- de-1901: German, traditional spelling
- de-1996: German, reformed spelling
- ja-latn: Japanese in Latin script
- ja-latn-hepburn: Hepburn spelling
- ja-latn-hepburn-heploc: Library of Congress method
- ase: American Sign Language
- sgn: Sign Language

Who needs a language tag like ja-latn-hepburn-heploc?

- Imagine a (small, but well sold) web application for
  - Automatic and / or manual subtitling Japanese videos on YouTube
  - Making them available for Japanese language learners
- The possibility of creating this cheaply is strongly connected to progress in other area of Web technologies, e.g. HTML5 video accessibility

Application in HTML5 <video> element (still in draft status)

```
<video src="http://example.com/video.ogv" ...>
 <text role="SUB" lang="ja-latn-hepburn-heploc"
    type="application/smil" src="japanese-lhh.smil"/>
 <text role="SUB" lang="ja" type="text/x-srt"
    src="translation_webservice/ja/caption.srt"/>
</video>
```

For details see
http://www.w3.org/html/wg/wiki/MultimediaAccessibilty

Language tags leads to layout ...

- Already possible for a while: selection of culture-specific glyph images based on language tags

```
<span xml:lang="zh-CN">[雪 zh-CN]</span>
<span xml:lang="ja">[ 雪 ja]</span>
<span xml:lang="ko">[ 雪 ko]</span>
```

[雪 zh-CN] [雪 ja] [雪 ko]

With new markets – reading electronic books in "Japanese Layout"
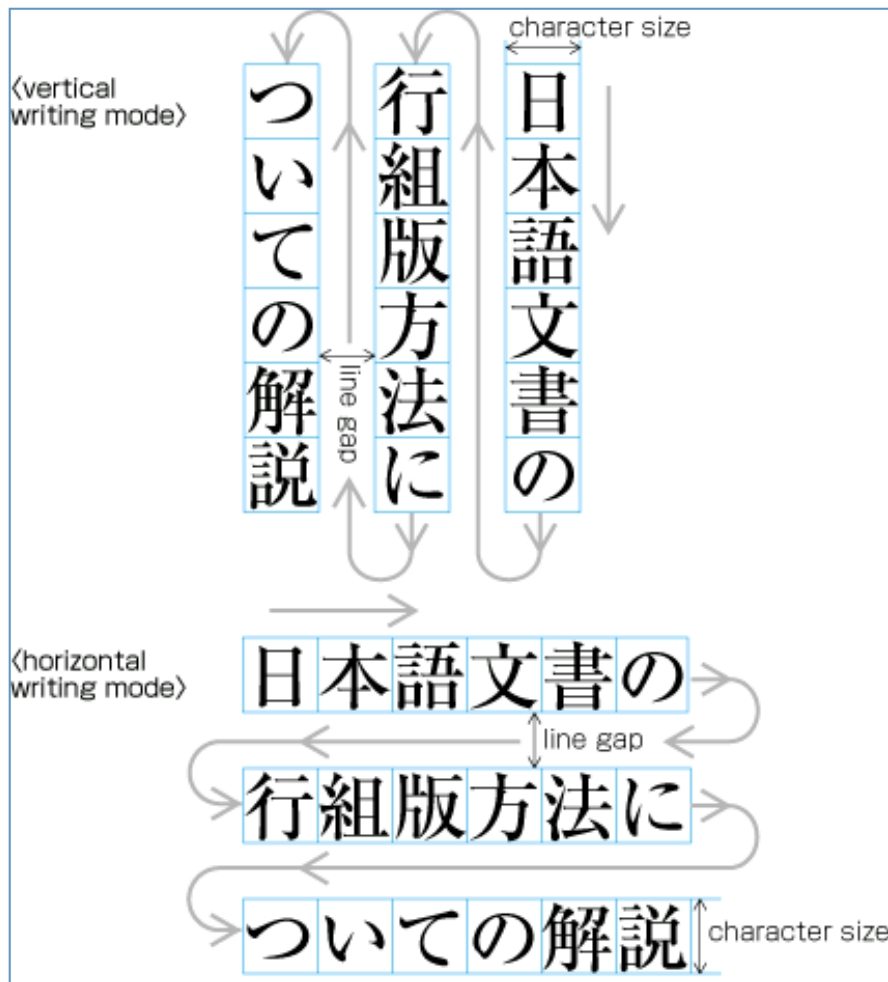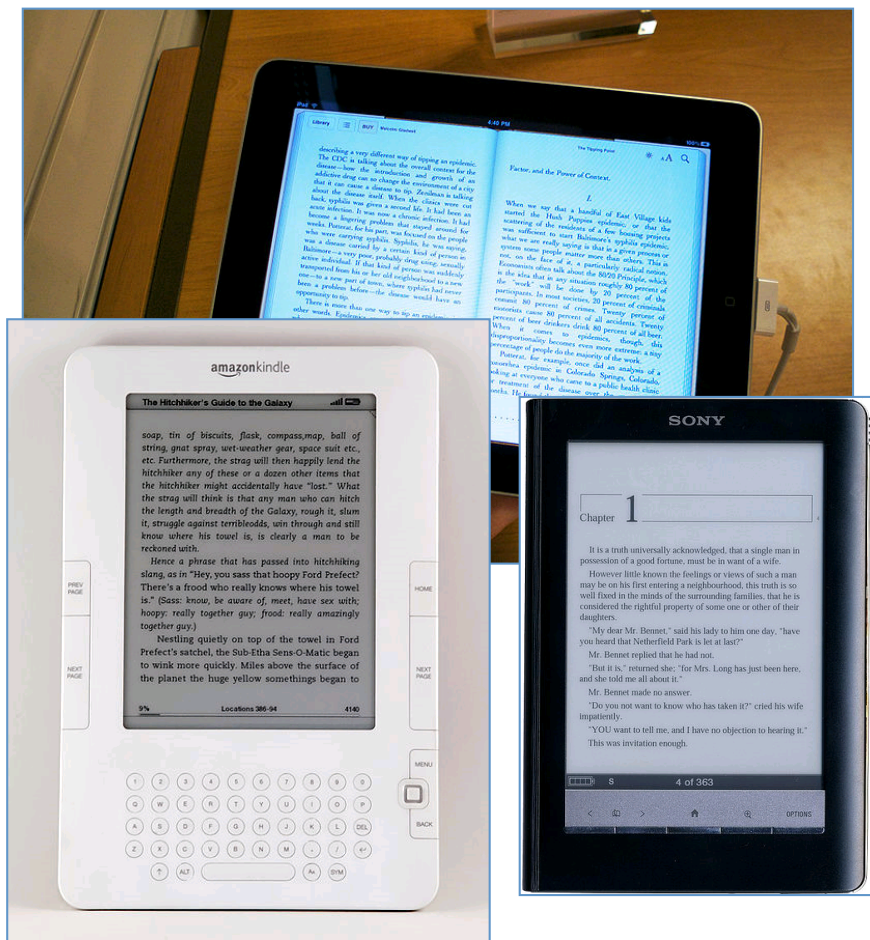
Details on Japanese Layout: see http://www.w3.org/TR/jlreq/

The tricky bits …

- An Example: Vertical layout = the same content, just set horizontally? No!
- E.g. marks for "emphasis": different characters depending on horizontal vs. vertical writing mode

Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- Projects "Multilingual Web" and "META-NET"

Bridging technology gaps

- Technologies in Web industry
  - HTML, XML, HTTP, …

- Technologies in the localization industry
  - Example TMX: Translation Memory Exchange

  - …

- How should these be combined for fast & <u>cheap</u> localization of content and services for many devices?

Metadata for the rescue!

Metadata 1: XLIFF

- XML Localization Interchange File Format
- Open OASIS Standard for representation of
  - Content to be localized
  - Metadata about the localization process

Basic Example: XLIFF File

- XLIFF files keep source and target content together

```
<trans-unit id="n1">
 <source>This is a sentence.</source>
 <target xml:lang="fr">Translation of "This is a sentence."
 </target>
</trans-unit>
```

Basic Example: XLIFF File

- Metadata in XLIFF files can help to integrate different means of translation (e.g. human vs. machine)

```
<trans-unit id="n1">
 <source>This is a sentence.</source>
 <target xml:lang="fr">Translation of "This is a sentence."</target>
</trans-unit>
<alt-trans match-quality="100%" tool="TM_System">
 <source>This is a sentence.</source>
 <target xml:lang="fr">TM match for "This is a sentence."</target>
</alt-trans>
```

Metadata 2: Internationalization Tag Set (ITS) 1.0

- W3C-Standard for Internationalization and localization of XML
  - Widely used (HTML, DocBook, DITA, ...) or special purpose formats
  - Content authored in these formats
  - An entry point to the localization tool chain
  - Adopted in localization industry tools like SDL Trados

Basic Principles of ITS 1.0

- Say important things
- About specific content
- In a standard way

Basic Principles of ITS 1.0

- Say important things: Do not translate

- About specific content: all "uitext" elements

- In a standard way: its:translate="no"

Two approaches for expressing the same information

```
<para>
    Press the
    <uitext its:translate="no">START</uitext>
    button to sound the horn. The
    <uitext its:translate="no">MAKE-READY/ RUN</uitext>
    indicator flashes.
</para>
```
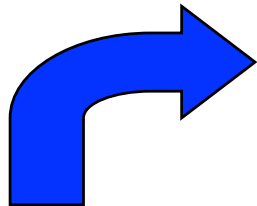
```
<para>
    Press the
    <uitext>START</uitext>
    button to sound the horn. The
    <uitext>MAKE-READY/ RUN</uitext>
    indicator flashes.
</para>
```

Global Approach

```
<its:rules ... its:version="1.0">
    <its:translateRule selector="//uitext" translate="no"/>
</its:rules>
```
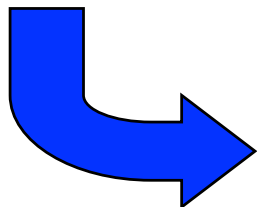
ITS 1.0 "Data Categories"

- Translate : Whether the content of an element or attribute should be translated or not

- Terminology: Mark terms and optionally associate them with information, such as definitions

- Directionality: Specify the base writing direction of blocks, embeddings and overrides for the Unicode bidirectional algorithm

- Ruby: Provide a short annotation of an associated base text, particularly useful for East Asian languages

- Language Information: Express the language of a given piece of content

- Localization Note: Communicate notes to localizers about a particular item of content

- Elements Within Text: Identify how an element behaves relative to its surrounding text, e.g. for text segmentation purposes

Why is this important? Example (1) where I18N / l10N Metadata helps in "long tail" localization

Language?

Terminology?

Codes?

Footnotes?

Foreign language expressions?

Annotations for readers?

```
<Собирание версия="1.2-3">
 <Объект id="12„>      <НомерОбъекта>OnlineCard</
НомерОбъекта>
   <ВНаличии>123</ВНаличии>
   <Описание xml:lang="ja">第二発電機</Описание>
 </Объект>
<Объект id="64">
 <НомерОбъекта>45-7894-456</НомерОбъекта>
 <ВНаличии>Latest Offer</ВНаличии>
 <Оп xml:lang="ja">手動ウォーター・ポンプ </Оп>
</Объект>
</Собирание>
```

Example (2) where knowledge about I18N markup usage helps in "long tail" localization

Volcanic eruptions have literally devastated large inhabited areas. During the 1914 eruption of Sakurajima in Kyushu, 687 houses in Kurokami were buried in hot ash. What remained of this shrine gate, previously five meters tall, was left as a reminder.

Kurokami maibutsu gate (腹五社神社黒神埋没鳥居), Sakurajima Island.

Suitable Markup?

```
<image src="kk-torii.jpg" height="180" width="240" caption="Kurokami maibutsu gate (腹五社神社黒神埋 没鳥居), Sakurajima Island." />
```

Better:

```
<image src="kk-torii.jpg" height="180" width="240">
  <caption>
  Kurokami maibutsu gate (
  <span xml:lang="ja">腹五社神社黒神埋没鳥居</span>
  ), Sakurajima Island.
  </caption>
</image>
```

Adopted from Richard Ishida (W3C)

XLIFF and ITS 1.0

- ITS 1.0: entry point in the localization chain

- A pre-requisite for properly internationalized vocabularies and content

- XLIFF: the meat of the localization chain

- ITS 1.0 makes XLIFF creation and processing easier

XLIFF and ITS 1.0 example: ITS2XLIFF

- See http://fabday.fh-potsdam.de/~sasaki/its/

- XSLT-based round tripping tool for generation of XLIFF from XML with ITS markup, and integration of translated content into the original XML

- Open source standard technology based - one example of how "long tail" content localization can become easier

Future usage scenarios for ITS 1.0: "Long tail" L10N via the Web
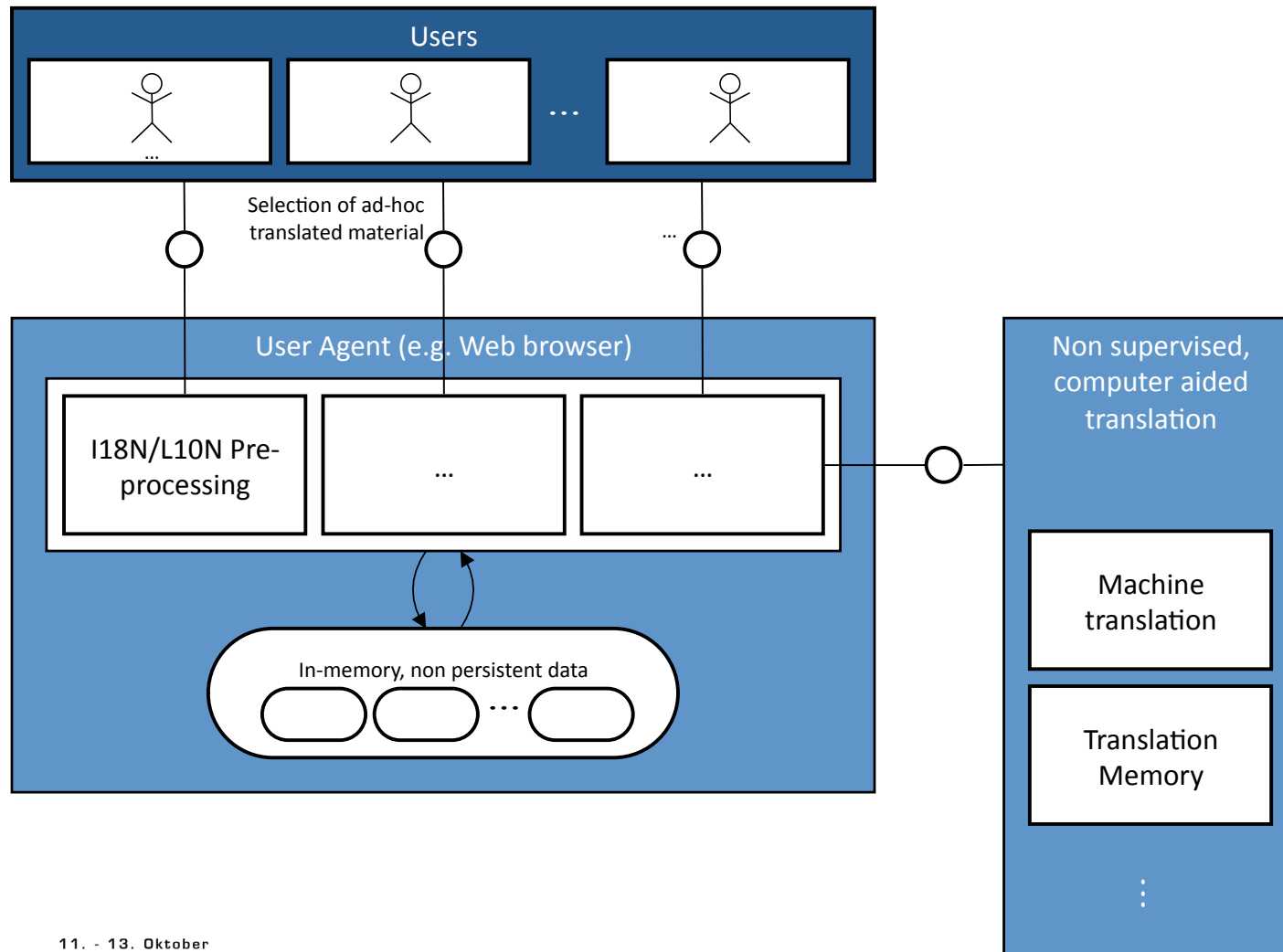
Who can help to make that happen?

- Of course you!

- If content authors & developers don't use ITS, tools which want to generate XLIFF from let's say HTML will produce a lot of crap

- We need you to fix that!

webtech
CONFERENCE 2010

11. - 13. Oktober
Rheingoldhalle, Mainz

Why we need you for the multilingual web

- Input from www.postbank.de

  „Ob Postbank direkt, Online-Banking, Online-Brokerage oder myBHW. Die häufigsten Fragen zu unseren Transaktionssystemen finden Sie an dieser Stelle."

- Output via Google translate

  "Whether Postbank direct, online banking, online brokerage or myBHW. Frequently asked questions about our transaction systems can be found at this location."

Why we need you for the multilingual web

- Input from www.postbank.de

  „Ob Postbank direkt, Online-Banking, Online-Brokerage oder myBHW. Die häufigsten Fragen zu unseren Transaktionssystemen finden Sie an dieser Stelle."

- Output via Google translate

  "Whether Postbank direct, online banking, online brokerage or myBHW. Frequently asked questions about our transaction systems can be found at this location."

Fixed terminology should not have been translated. If a content author / editor / developer (= you too!) would have used ITS "translate", Google translate would have worked.

What do we learn from this? Automatic language processing is better ...

- ... if you help it with metadata
  - About what is translatable or not
  - About terminology
  - About segmentation ("What is a footnote?")
  - ...
- Your employer will like it too
  - (Long tail) localization with high quality metadata gets better
  - The whole localization process gets cheaper ☺

The question is now: Who will start?

- Content creators don't use metadata since nobody processes it

- Nobody processes metadata since there is none

- The first company who breaks that vicious circle and develops a successfully model of deploying metadata (see previous slides) will get rich ☺ ☺ ☺

The question is now: Who will start?

- Some people already have started: WAI-ARIA (cf. talk from Tomas Caspers at this conference)
  - Metadata for "roles" of specified elements
    - "I am navigation"
    - "I am a tab"
    - …
- Used e.g. by screen readers
- Shows common aspect of barrier free web design and web design for the "automatable, multilingual web": appropriate metadata

Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- Projects "Multilingual Web" and "META-NET"

See http://www.multilingualweb.eu/
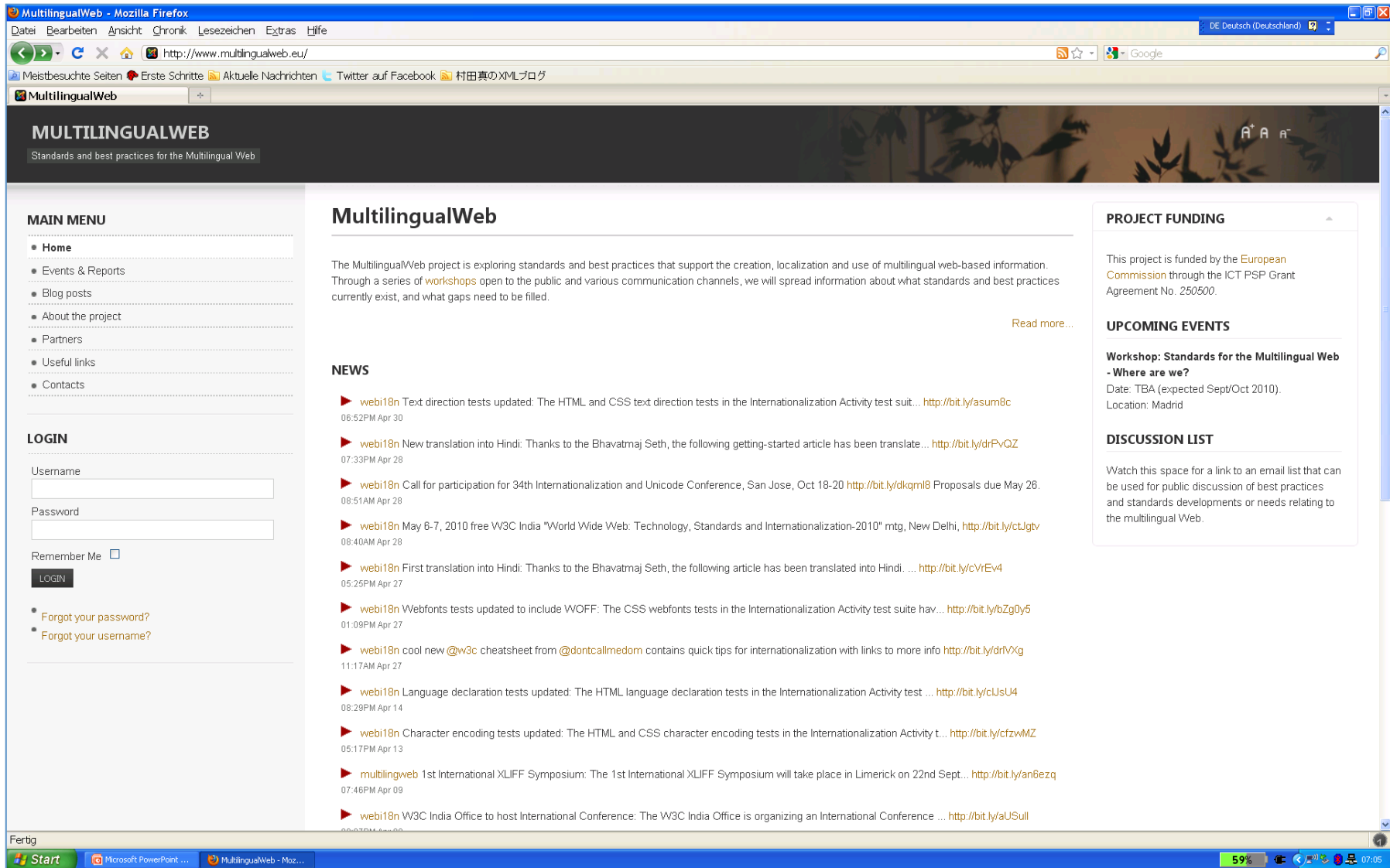
Background

- Participants from industry and academia (e.g. computational linguistics)
- Aim: Bridge the gaps between industries (and research areas) described
- Education about new developments (e.g. in the area of language tags, layout, web based localization) – like in this presentation ☺
- More mutual understanding of users needs and what tools can do
- Support via special purpose tool
  - First example "I18n checker" http://rishida.net/tools/i18nchecker/

Participants from: Web, Localization industry, (automatic) translation research, ...

- <u>ERCIM/W3C:</u> coordination
- CNR-ILC
- Facebook Ireland
- The University of Applied Sciences (UAS) Potsdam
- Institut Josef Stefan (JSI)
- Institutul de Cercetari Pentru Intelegentia Articificiala (RACAI)
- The Language Technology Centre
- Lionbridge Belgium
- Microsoft Ireland
- Opera Software

- SAP
- The Translation Automation User Society (TAUS)
- Teknillinen Korkeakoulu
- University of Oviedo (ILTO)
- Universidad Politécnica de Madrid (UPM)
- The Language Resource Centre
- University of Economics, Prague
- Transware Ltd (WeLocalize)
- XML-INTL

Workshops as a means for community building - topics

- The landscape of multilingual Web standards and best practices

- Authoring of the Multilingual Web

- Translation tool support (with focus on standards like ITS 1.0, XLIFF, TMX)

- Further topics – to be decided
  - Your input is more than welcome!

1st Workshop: "The Multilingual Web – Where Are We?"

- 26-27th October, Madrid

- Goal: Bring developers (you!), content creators, localizers, users, machine processing folks and policy makers together

- Provide input for upcoming work

- Program details at

www.w3.org/International/multilingualweb/madrid/program

META-NET

- EU-funded project, closely related to "Multilingual Web"

- Main aim: build an alliance for improving language technologies in Europe

- Laaarge: soon 40+ participating organizations in 30+ countries

- Very important: bring users of language technology in

META-NET

- Users and language technology companies = in Europe not only large companies, but more and more small SMEs

- Target of META-NET are these small and fast units – including you ☺

- EU has started special funding programs for SMEs – see http://tinyurl.com/eu-lt-sme ("objective 4.1")

META-NET

- Event: META-NET Forum
- Brussels, November 17th/18th
- Aim: Bring users / language technology developers / policy makers together
- Discuss a road map for the next 10 years of language technology road map and its applications
- Details and registration at

http://www.meta-net.eu/events

Overview

- Multilingual Web – what's that?
- I18N and L10N – traditional topics
- "The long tail" and its consequences for the multilingual web
- I18N and L10N on the Web - revisited
  - Traditional topics & new aspects
  - New: bridging technology and market gaps
- Projects "Multilingual Web" and "META-NET"

webtech CONFERENCE 2010

11. - 13. Oktober
Rheingoldhalle, Mainz

Summary

- The basic infrastructure of the multilingual Web is set
- Nevertheless, even core parts like language identification have are being updated to easy international use for many & diverse audiences
- Localization is getting diverse across industries (Web, Localization, including automatic translation)
- Specific (Meta)data formats like XLIFF and ITS 1.0 can help bridging gaps between players and users in the field
- We need you as a metadata user to realize faster and cheaper localization
- The "Multilingual Web" project aims at being a place for gathering possibilities for a "long tail" multilingual web – a web for really everyone's business!
- The "META-NET" project aims at building an alliance between language technology providers and users to make the long tail localization model happen

Thank you for your attention!
Vielen Dank für Ihre Aufmerksamkeit
ありがとうございました