Internationalisierung bei XML

Felix Sasaki

DFKI / Fachhochschule Potsdam

W3C deutsch-österr. Büro

felix.sasaki@dfki.de

Markupforum 2011

Über mich

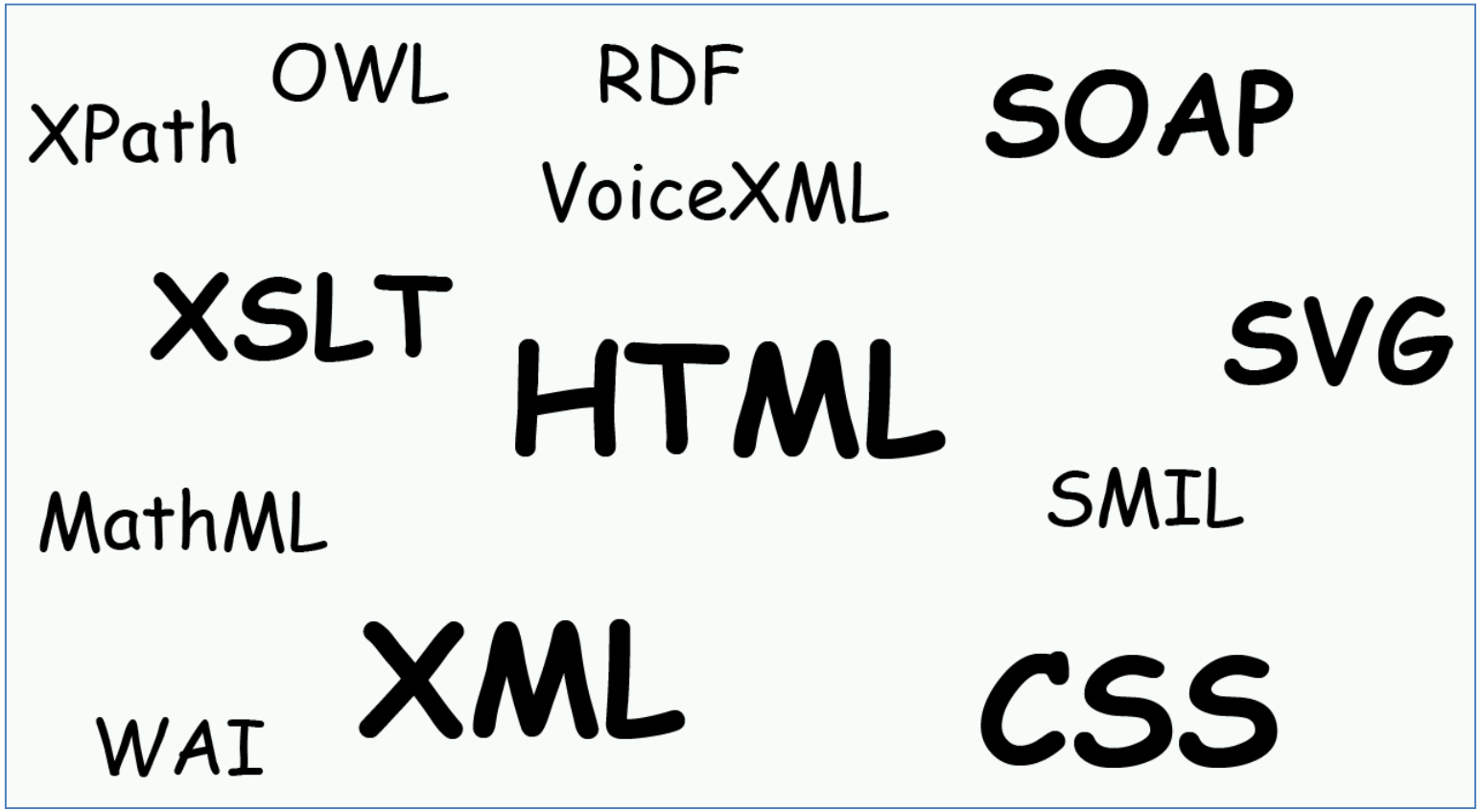
- Studium der Japanologie und Linguistik in Deutschland und Japan
- Dissertation im Bereich Computerlinguistik zu Webtechnologien und mehrsprachigen Daten
- 2005-2009: Arbeit in Japan beim W3C, hauptsächlich in der „Internationalization Activity“
- Seit 2009: Professor an der FH Potsdam / Manager des W3C deutsch-österr. Büro
- Seit Herbst 2010: Senior Researcher am DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz)

Über W3C Büros

- Ein Kontaktpunkt wenn man ...
  - W3C (noch nicht) gut kennt
  - Spezifische Fragen hat wie „Wer arbeitet an Thema ABC ...“
  - Neue Themen in Webstandardisierung einbringen will und sich fragt wo sie passen könnten
- Bitte sprechen Sie uns an – zu obigen Themen, und sonst auch ☺

INTERNATIONALISIERUNG BEI XML – EINIGE TRADITIONELLE THEMEN

Nutzung von Unicode in (XML)-Technologien



Internationalisierte Webadressen

- Internationalized Resource Identifier (IRI)
- I18N im Pfad einer Webadresse, z.B.:

Sprachenidentifikation via xml:lang

Sprachtags wie “en”, “en-us”, “de”, “ja”, …

Für

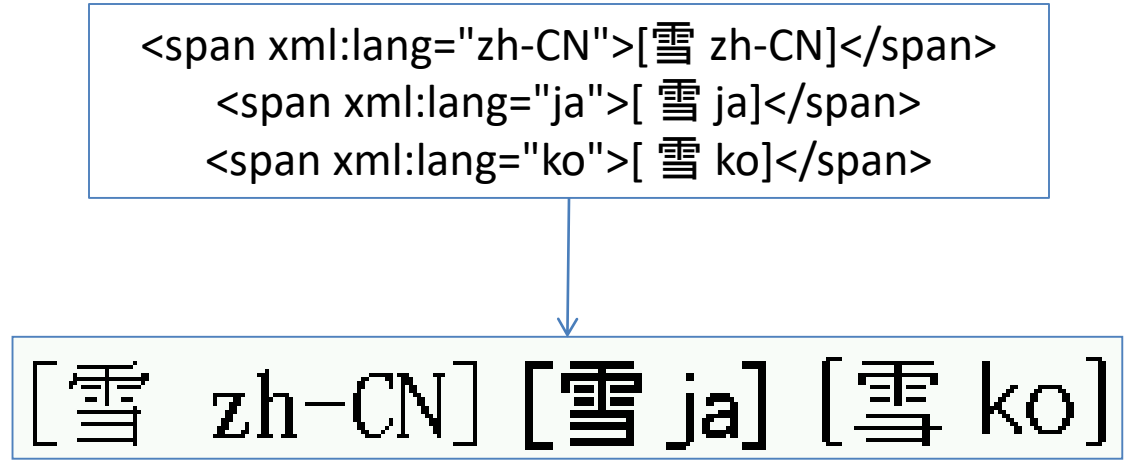
- content-negotation

Sprachenidentifikation via xml:lang

Sprachtags wie “en”, “en-us”, “de”, “ja”, ...

Für

- content-negotation
- Sprachspezifisches Layout

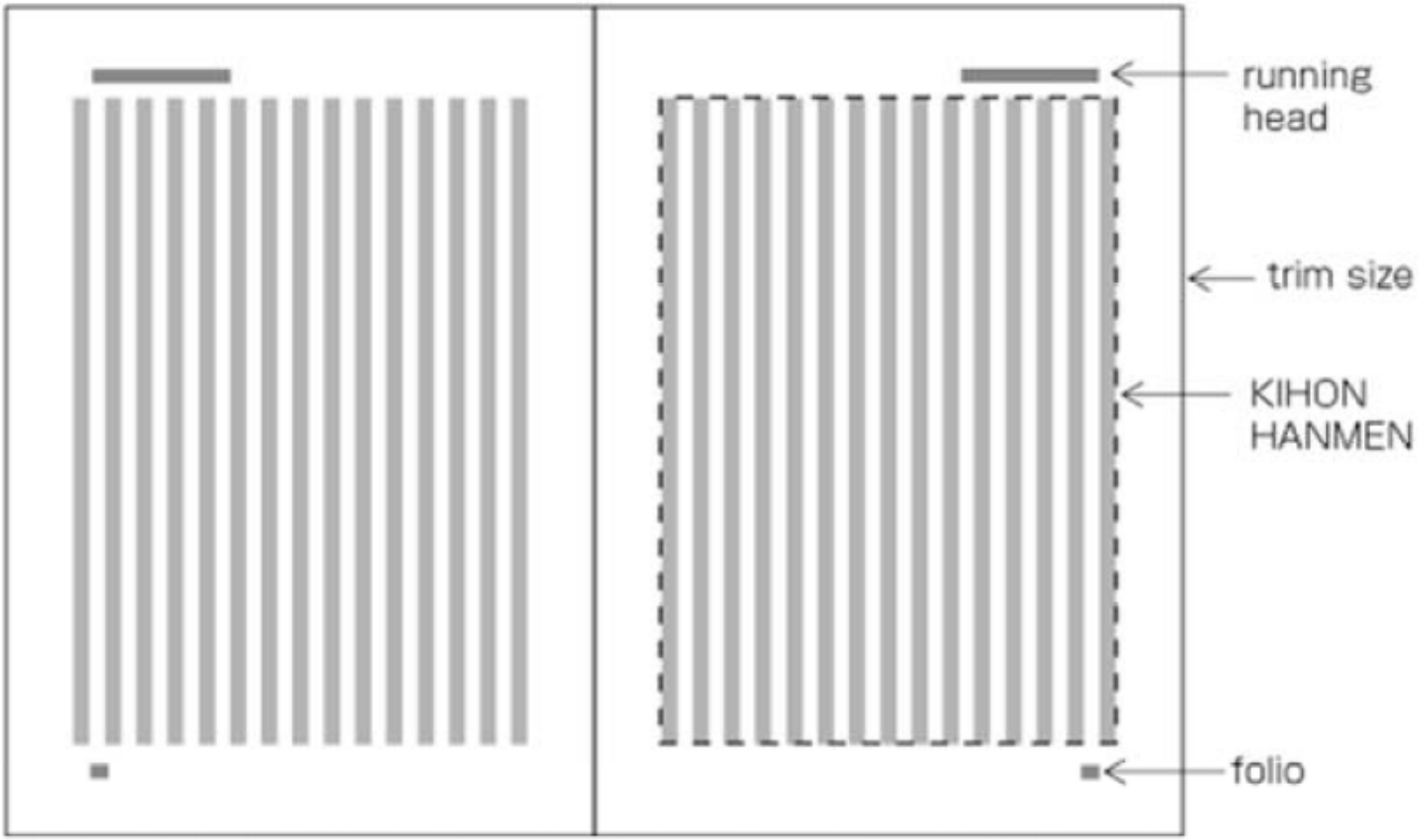


NEUE THEMEN I: KULTURSPEZIFISCHES LAYOUT AM BEISPIEL „JAPANISCH“

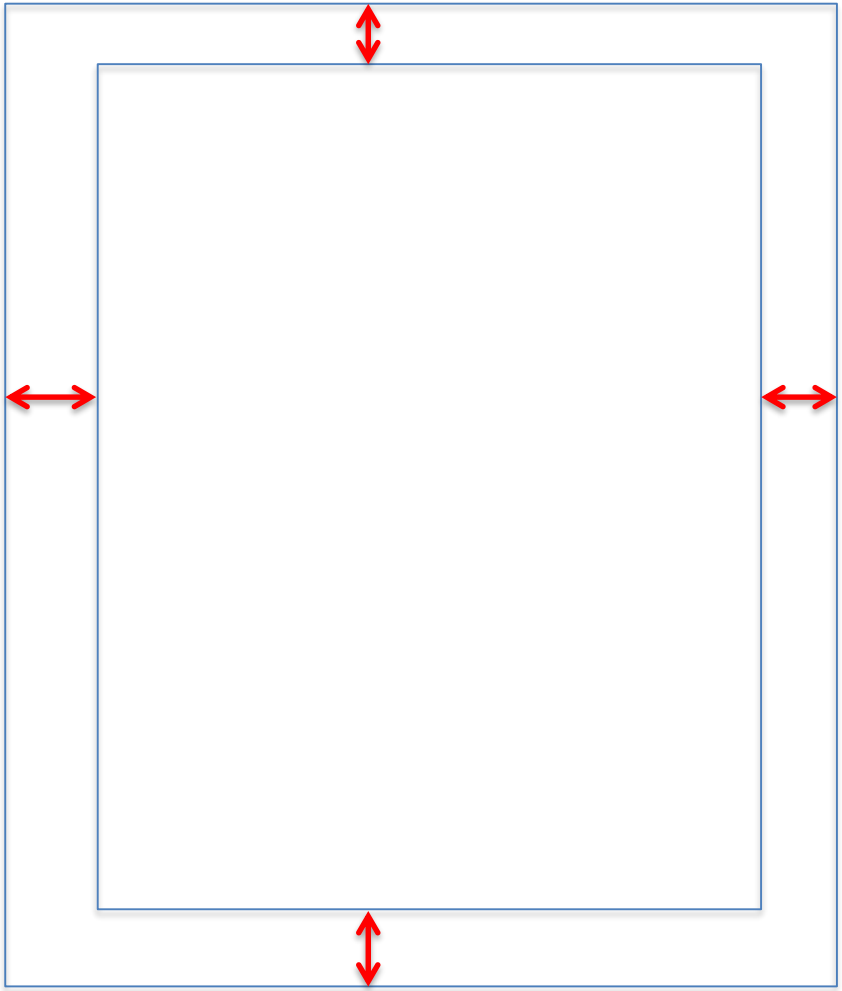
Japanisch

- Geschrieben in vier Schreibsystemen
  - Kanji (漢字)
    - Basiert auf chinesischen ideographischen Zeichen
  - Hiragana (ひらがな)
    - Lautschrift für japanische Wörter
  - Katakana (カタカナ)
    - Lautschrift für ausländische Wörter
  - Romaji (romaji)
    - Lateinisches Alphabet

Neue Layout-Konzepte: Beispiel KIHONHANMEN

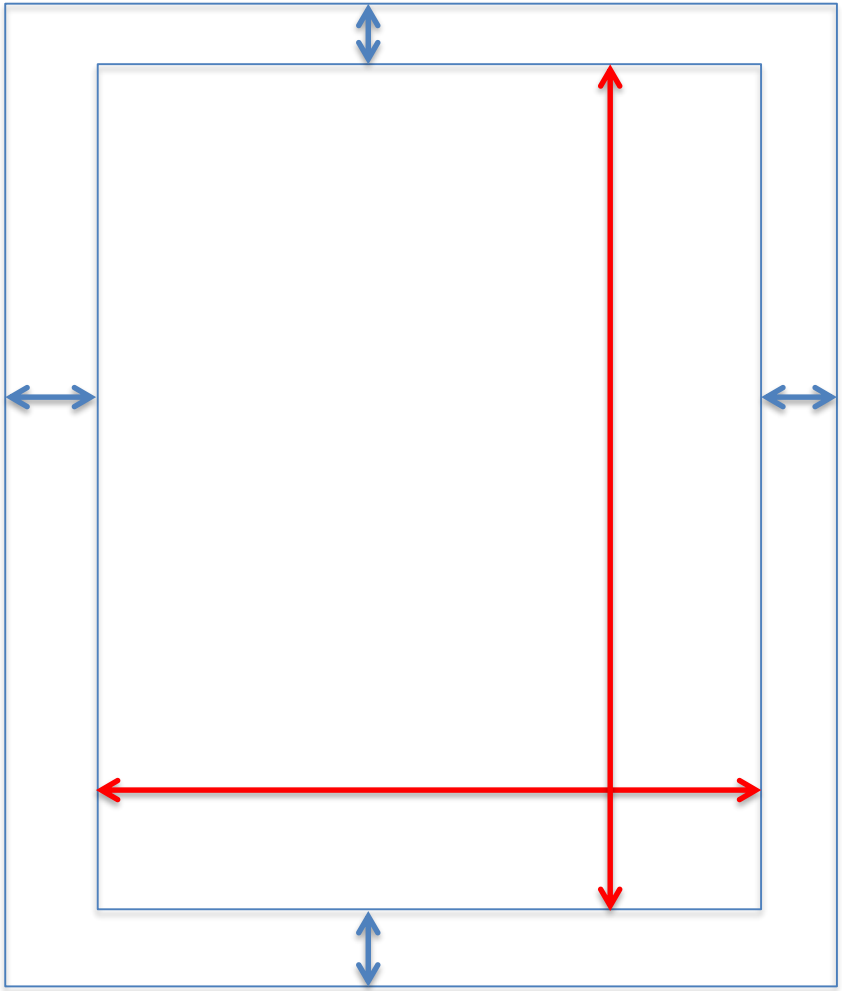


„Westliches“ Seitenlayout



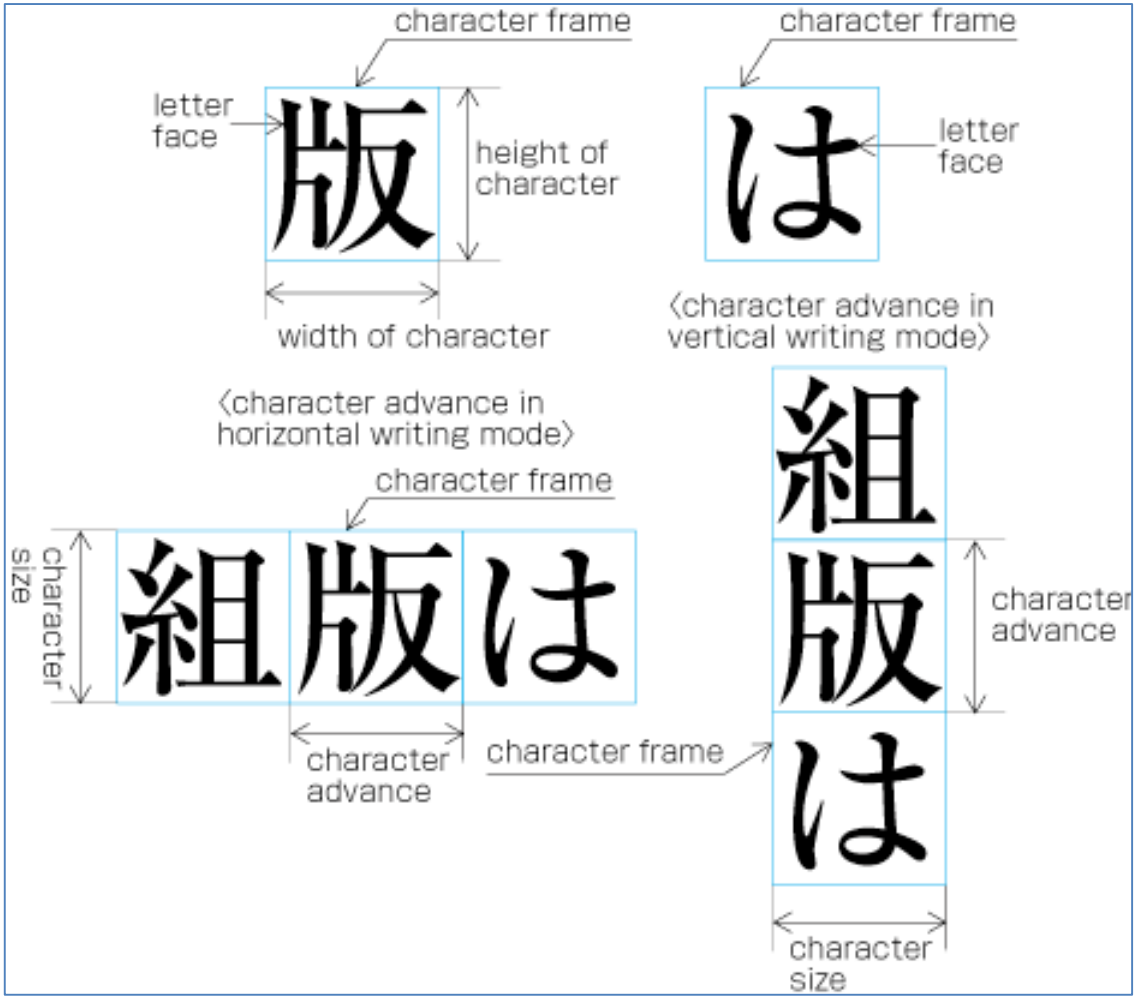
- Zunächst Festlegung der Seitenränder
- Dann Festlegung des Druckbereichs

Japanisches Seitenlayout



- Zunächst Festlegung des KIHONHANMEN anhand von Zeichengröße, Zeichenzahl, Spaltenzahl, Spaltenabstand
- Dann Festlegung der Seitenränder

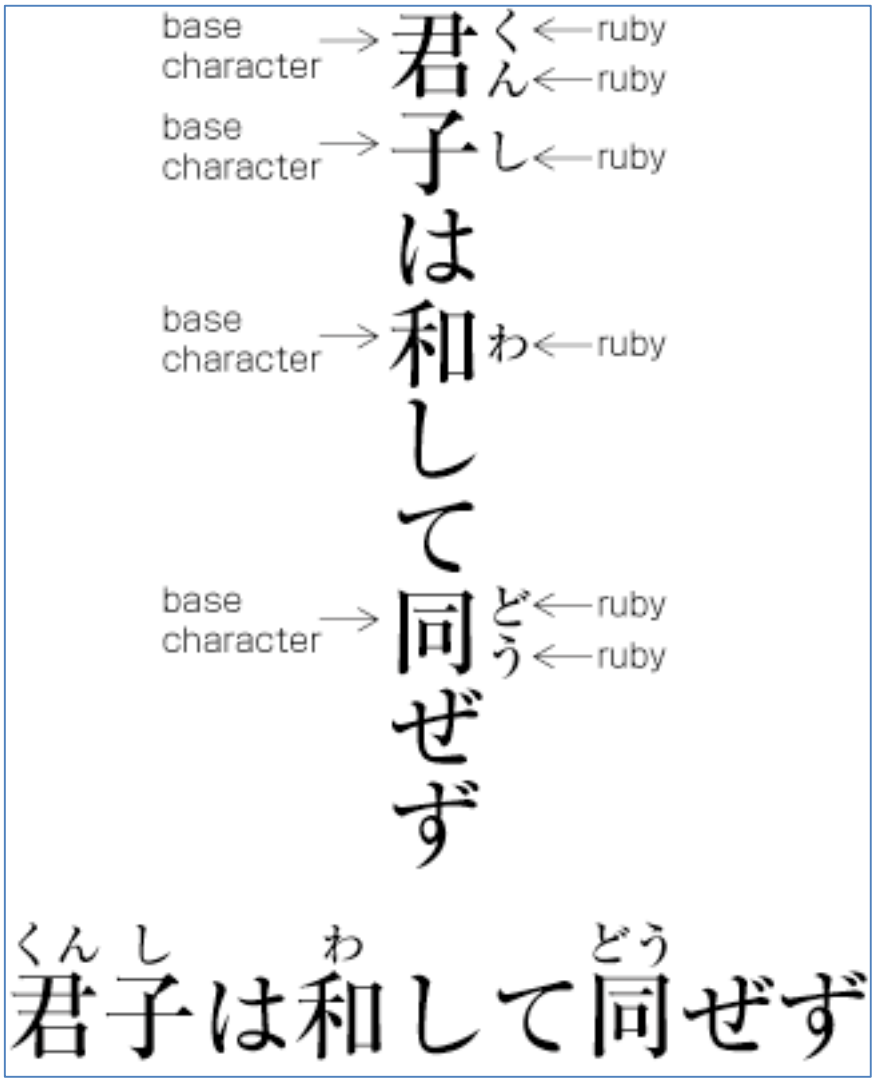
Hintergrund: reguläre Ausmaße japanischer Zeichen



- Vgl. Dokument „Requirements for Japanese Text Layout“

http://www.w3.org/TR/jlreq/

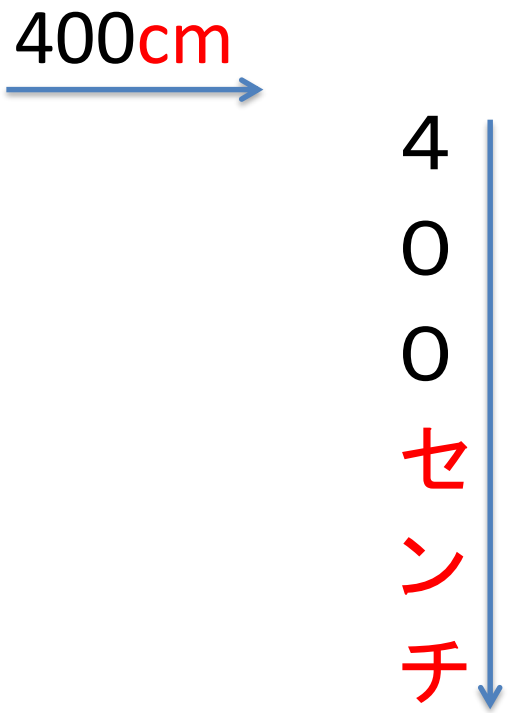
Neue Layout-Bestandteile: Ruby



- Annotationen zum Basistext
  - Aussprache, Erklärung, …
- Darstellung via sprachspezifischer Layoutregeln

Herausforderung für „Single-source Publishing“

- Andere Inhalte in Abhängigkeit von der Layoutrichtung



Zusammenfassung zu kulturspezifischem Layout

- Neue
  - Konzepte (vgl. KIHONHANMEN)
  - Bestandteile (vgl. Ruby)
  - Sichten auf Verfahren (z.B. „Single-source Publishing“)

= neue Terminologie!

- Wo kommt XML?

Input und Output: „W3C Japanese Layout Task force“

- Teilnehmer: Experten der japanischen Druckindustrie + aus allen Layout-relevanten Arbeitsgruppen
  - CSS
  - XSL
  - SVG
- Ähnliche Gruppen im W3C für Layout im Chinesischen und Koreanischen
- Einfluss auch auf die Entwicklung von ePub 3.0

NEUE THEMEN II: MEHRSPRACHIGKEIT

Internationalisierung:

- Basis (Zeichenkodierung, Sprachenidentifikation)
- Erweiterungen hinsichtlich Darstellung (internationales Layout)
- Erweiterungen hinsichtlich Informationsverarbeitung: Mehrsprachigkeit (mit automatischen Mitteln)
  - Automatische Übersetzung, Zusammenfassung, Qualitätskontrolle, ...

Was man für Mehrsprachigkeit im Web braucht

- Input von www.postbank.de

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

- Ausgabe 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.”

Lücke 1: Maschinen nutzen keine Metadaten in der Eingabe

- Input von www.postbank.de

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

- Ausgabe 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.”

Feste Terminology

Sollte nicht übersetzt werden. Wenn ein Autor diese Information markiert hätte, wäre das automatische Tool besser

Lücke 2: Maschinen kennen keine Prozesse zur Datenerzeugung

- Input aus einer Datenbank – dem „hidden web“:

  „Ob <term>Postbank direkt</term>,
  <term>Online-Banking</term>,
  <term>Online-Brokerage</term> …“

- Ausgabe im Web:

  „Ob Postbank direkt,
  Online-Banking,
  Online-Brokerage …“

Feste Terminologie (= Metadaten) …

Publikations-prozess

… wird verloren im Web ☹

Lücke 3: keine eindeutige Identifikation

- Von Metadaten und Verarbeitungsprozessen (vorherige Folien)
- Von Ressourcen – was ist z.B. ein Lexikon
  - In maschineller Übersetzung?
  - In Lokalisierung?
  - Für den menschlichen Leser?
  - …
- Wiederverwendung und Kombination von Ressourcen wird behindert

Wer kann diese Lücken füllen?

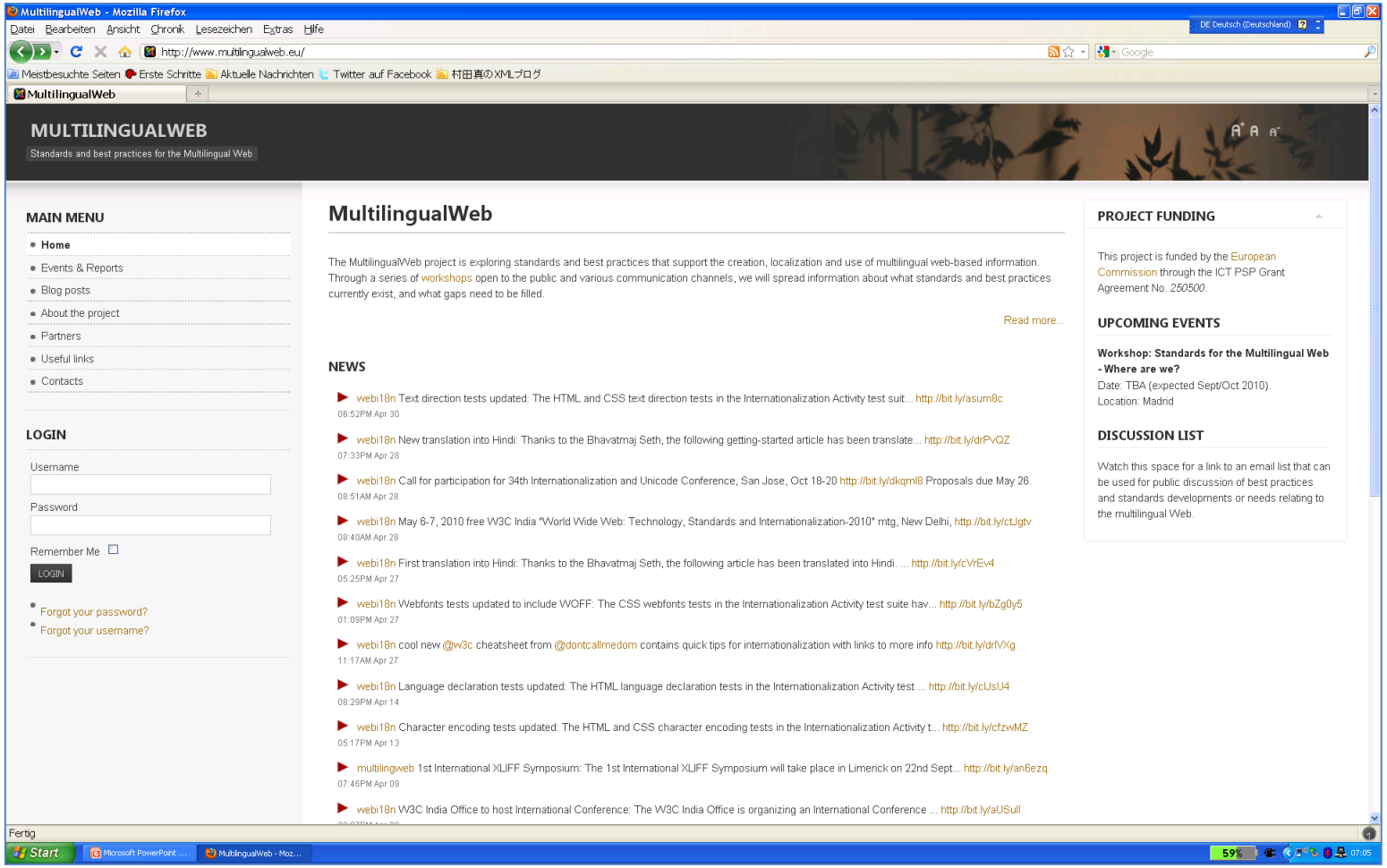
- Autoren(systeme)
  - Nutzung von Terminologie / Markierung von Übersetzbarkeit z.B. in CMS
- Lokalisierer
  - Lokalisierungsworkflows sensibel für Metadaten (Quelltext- und Prozess-bezogen) machen
- „Sprachtechnologie" Experten
  - Tools sensibel machen für Metadaten im Quelltext und im Prozess
  - Ressourcen und Workflows klar beschreiben

Wie können die Lücken gefüllt werden?

- Metadaten standardisieren
- Metadatennutzung propagieren bei
  - Erzeugern von Inhalten
  - Verschiedenen Gliedern der Verarbeitungskette
- Anwendungsszenarien Community-übergreifend definieren

ZUM SCHLUSS: PROJEKTHINTERGRUND

EU-Projekt „Multilingual Web“



Vgl. http://www.multilingualweb.eu/

Hintergrund

- Teilnehmer aus Industrie und Akademia (z.B. Computerlinguistik)
- Ziel: Lücken zwischen Industrien, Nutzern und Forschern schließen
- Outreach zu neuen Entwicklungen (z.B. hinsichtlich internationalisiertem Layout) – wie in dieser Präsentation ☺
- Mehr gegenwärtiges Verständnis für Bedürfnisse von Nutzern und Möglichkeiten (automatischer) Verarbeitung)
- Toolentwicklung
  - Beispiel “I18n checker” http://rishida.net/tools/i18nchecker/

Teilnehmer

- ERCIM/W3C: 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 zum Community-Bildung

- Erster Workshop 26.-27. Oktober 2010, Madrid: „The Multilingual Web – Where Are We?“
- Zweiter Workshop 4.-5. April 2011, Pisa: „Content On The Multilingual Web“

EU-Projekt „META-NET“

- Enge Verbindung zu „Multilingual Web“
- Hauptziel: Langfristige Allianz für Sprachtechnologie in Europa bauen
- Umfasst mehr als 40 teilnehmende Organisationen aus 30+ Ländern
- Wichtig: Nutzer von Sprachtechnologie involvieren

META-NET

- Nutzer und Sprachtechnologiefirmen = in Europa oft KMUs
- Ziel von META-NET sind schnelle und flexible Einheiten – wie Sie ☺
- Die EU hat entsprechende Förderprogramme veröffentlicht - vgl. http://tinyurl.com/eu-lt-sme („objective 4.1“)

META-NET

- Event: META-FORUM 2011
- Budapest, 27.-29. Juni 2011
- Ziel: Nutzer / Sprachtechnologieentwickler / „Entscheider“ zusammenbringen
- Ziele für die Sprachtechnologie in den nächsten 10 Jahren diskutieren
- Details und bald Registrierung unter

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

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