Parlez-vous zwo-null, señor?
Providing Cross-National and Multi-Lingual Web Communities

October 22th, 2008
Andreas Ravn
Senior Software Engineer
namics (deutschland) gmbh
Hamburg
Worth ladies and Mr.,
I be pleased that nevertheless or other one grants too to this early that away was not afraid and to this lecture with the title parlez vous two zero senor to appear here.
With one another will we in the next grants a hopefully completely stimulating trip the world from chances and tücken multilingual Web of communities to undertake and now geht's also equivalent loosely.
Parlez-vous zwo-null, señor?
Providing Cross-National and Multi-Lingual Web Communities

October 22th, 2008
Andreas Ravn
Senior Software Engineer
namics (deutschland) gmbh
Hamburg
Web 1.0
Web 1.0

» Content centrally created by editorial staff
» Internal Quality Assurance for content
» Content inheritance
» Control over
  — website structure
  — Languages used
  — Localized content
رائع

محتوى لطيف، مذاق

IKEA Food
نقانق 3 ريال

خططوا منازلكم حسب طريقة حياتكم

أفكار ملهمة لغرفة النوم
المؤملة تجاه البيئة
سلامة الأطفال على طريقة إيكيا
مبادئ التأثير، فن تناول الطعام
acreli المزيد

جميل غرف النوم
وعيدنا هو...
اعرف أكثر
امرحوا. استمتعوا بوقتكم، لتأخذ "فيا"
Web 2.0 applications (and the like)

» social networks
» communities
» content portals
» grassroot
» aggregation sites (lists, social bookmarking)
» crowdsourcing sites
» discussion groups
» weblogs
» wikis

...
In Web 2.0

» User generated content
  – content elements
  – reviews
  – comments
  – profiles
  – discussions

» User generated structure
  – tags
  – lists
  – links
  – rating, reputation
Where does your content come from?

- Social networks
- Centrally generated
- User-generated content and structure
- Communities
- Content portals
- Forums
- Aggregation
- Wikis
- Weblogs

Centrally generated

User-generated
Web 2.0

» Provider holds less control over
  - Content creation
  - Categorization of content
  - Quality of content

» How to organize a multi-lingual community?

» How to organize a community that has members in different countries?
Why bother?
Why mix content and users in the first place?

» more members = more content, more exchange
» maybe higher quality level
» community model may require multi-country context (or benefit from)
» Yes, we’re international!
Now what’s so difficult about this?
« Localization is a technical problem and will be solved by my developers in no time. »

— a customer
(localization not online yet)
Localization issues.

» Technical
» Content
» Organizational
» Cultural
» Marketing
Issue: Pinpoint the user.

» Where does my user come from and what languages does he speak?
  - browser language
  - chosen domain
  - IP address tables
  - profiles

» If wrong, the user knows better, and can configure.
Issue: Encoding, character sets, codepages

» Special characters
  - iso-8859-1 = latin
  - iso-8859-5 = cyrillic
  - iso-8859-9 = greek
  ...

» Solution (mostly): UTF-8
  - Now standard; generally not a problem anymore
  - Check software components
Issue: Content and structure

» User Generated Content, sure.
» Structure: e.g. tags
Can we translate user generated content?

» Automatically
  – Translation services exist
  – e.g. Google Language API (translation, language detection)
  – Problems:
    – General quality
    – special vocabulary
    – fixed terms, technical and functional terms
    – **UGC: You speak on behalf of the user!**
      – bad translation reduces the author’s credibility
Can we translate user generated content? (2)

» Automatically

» Editorial
  – Lots of work!
  – Featured user articles, focus on „best rated“

» Community
  – Involve community
  – Reward advertising
Localization Is Not Just Translation

» Simpler aspects: Date/Time Formats, Timezones, Currencies, Temperature, Measures

» Work & research: Audio, video

» Culture: Meaning, customs and taste (e.g. image types, colors) vary in different countries

» Don’t forget: legislative provisions, etc.
Issue: Searching (Full Text Search)

» What language are the search terms in?
   – Use context (language, profile, country) or language detection
   – Spelling

» How to display the result?
   – Tokenizing, stemming, similarity/proximity
   – Result set: what to show? Mix languages?
   – Result structure: relevance!
More issues.

» Geocoding: different standards
» Administration: e.g. user support
Some Approaches.
Approach 1

„Web 1.0“
1. „Web1.0“. 

» How it works.
   – Separate communities for every language and/or country.

» Advantages
   – No localization problems

» Disadvantages
   – No intl. web community at all
   – Redundancy
   – Critical user mass for each community
Approach 2

„Laissez-faire“
2. Laissez-faire.

» How it works.
   – The user is free to use any language he likes.
   – The frontend is localized, thus giving the impression of a “web 1.0” site.

» Advantages
   – Easy to implement
   – Better language level in the individual language

» Disadvantages
   – Navigation, search, categorization diffuse – result quality suffers
Approach 3

„Common Ground“
3. „Common Ground“

» How it works
  – Offer the community platform only in the “lingua franca”

» Advantages
  – Includes most of your users.

» Disadvantages
  – Most common language still leaves out large relevant groups (e.g. French also a large community)
  – Language level altogether lower (in foreign language)
Approach 4
„S.O.D. “
4. „S.O.D.“

» How it works.
  – Provider chooses the language to be used.
  – Frontend is available only in this language

» Advantages
  – Easy to implement

» Disadvantages
  – Foreign users are not encouraged to contribute content
Approach 5
„Babelfish“
5. „Babelfish“

» How it works.
   - Contributors use their own language to enter content
   - All content is presented in the user’s own language

» Advantages
   - Maximum coverage of the community
   - No need to learn Esperanto or Volapük! ;-)

» Disadvantages
   - Automatic translation necessary. Quality issues.
   - Cultural problems
What Others Do
How much localization do you need?
How much localization do you need?

» Check against your community needs
  – e.g.: local.ch in Switzerland
  – covers all of Switzerland, but
    – CH has strict(-ish) regional language zones
    – services are used mostly locally
Controlled vocabulary
Prefer „controlled vocabulary“. 

» Structured vs. unstructured data

» Avoid free text (where you can):
  – Text analysis is complicated and still nonsatisfying

» What are your „core assets“?

» What data can be made structured?
  – data that is language-neutral (e.g. Names)
  – Prefer pre-defined options to free text
    – naturally pre-defined (e.g. male/female)
    – definable by yourself (e.g. categorizations at ebay)
  – easily translatable and probably already available data (e.g. City names)
What data to show to the user?

» Be consistent.

» A good practice:
  — show all **structured** data in a translated form (e.g. in a user profile)
  — Let user enter his **spoken** languages in his profile (and show those)
  — offer to present other-language data
    – either in native language
    – or translated („Would you like to see…“)
    – shows awareness of multi-language problem
Gentle integration of other language/country content

» Priorize for **relevancy** (once again)

» You **may offer** automatic translations…
  - … but ask before doing so…
  - … and clearly label an automatic translation.
  - Check quality. Allow user feedback.

» Decide according to your community model, if country content can be mixed
  - also cultural issues
Internationalization from the beginning

» Every bit of content has an origin
  – For each tiny content element, save the language and region.

» Some software systems offer multilanguage support
  – e.g. Drupal: user indicates his home country, the primary language, other languages he speaks
  – Automatic composition of regionalized pages

» Use language-neutral page templating
  – Technical issue: Separate user interface and code, keep page templates language-neutral
Translation Files

MSG_NEW_MAIL_ALERT = „Good morning, {USER_FIRST_NAME}. You have {NUM_NEW_MAILS} new mails.“
Grammar and special formats

{USER_NAME} {START_DATE} {END_DATE} {TXT_IS_ON_VACATION}
„Andreas Ravn from 18.10.2008 to 21.10.2008 is on vacation.“
Welcome, Andreas!
Welcome, Mr. Ravn!
So what should I do?
Thank you.

Danke.
Mahalo.
Merci villmah.
Ookini arigatou.
Néá'êshehemeno.
Wokol a wala.
Dêkuji.
Bedankt.
Paljon kiitoksia.
Gracias.
Nagyon köszönöm.
Qujanarssuaq.
Rakux u kapamaxemaxes namen dimo.

Andreas Ravn

namics (deutschland) gmbh
andreas.ravn@namics.com
http://www.namics.com