MULTI-LINGUAL CONTENT MANAGEMENT

Jacob Palme

Stockholm University, Forum 100, 164 40 Kista, Sweden

e-mail:jpalme@dsv.su.se

Abstract

Some web sites provide their information in multiple languages. This paper discusses the experience in developing such a web site (http://web4health.info/). The development is greatly simplified, if the software clearly separates between language-independent structure and language-dependent information, so that changes in the language-independent structure can be done in one operation for all languages. Important is also work flow support, since different people do different tasks in the production, such as the writing of a text, its translation to English and its translation to other languages.

1 Introduction

1.1 Multi-lingual web sites

More and more often, organizations need to provide their information in several languages. Many web sites offer visitors a choice of which language to use in viewing the contents of the web site. The management of such a web site raises a number of issues. This paper reports on the experience from the actual development of such a web site.

1.2 The Web4Health web site

The web site developed has the name Web4Health, at the address http://web4health.info/. When this is written (May 2005), it contains about 800 informational texts for laymen in the area of mental health. Most of the content is available in German, English and Swedish, some of it also in Greek and Italian. In the month of April 2005 the web site had more than 230 000 visitors who viewed more than 600 000 pages.

The content of the web site was developed by medical experts in Germany, Greece, Italy, the Netherlands and Sweden. Each medical expert provided texts in their native language and/or English, and also translated the informational pages from other languages (mostly English) to their native language.

Most of the content is the same in all languages, but each medical partner was free to decide what to include and also could modify the text to suit the needs of each language

region when translation to his/her language, and was also free to add additional pages only available in a specific language.

The services provided by the web site for its visitors are:

1. Access to the informational pages through a hierarchical structure (taxonomy) of menus. One page can be placed in multiple positions in this structure.
2. Access to the informational pages using a natural-language question-answering system.
3. An ask-the-expert area, where visitors could ask a question not covered by the web site and get a personal answer from one of the medical experts.
4. Forums for discussion of mental health issues.

1.3 Content-management system

To manage the development and translation of the content, a multi-lingual content management system was developed for this project. Content management systems [11] are software systems specifically designed to handle large sets of documents, such as web sites with many pages [2]. According to [9], there are more than 225 software vendors supplying content management systems, even though this is a very new market, which has only existed for a few years, but our system has special features, described in this paper, not available in most other such systems (A system which has some such features is described in [10]). This paper describes the main principles of our content management system. It will explain the advantages of the system design, but also discuss drawbacks and how an ideal multi-lingual content management system should work. It also describes some features which we now understand that we should have implemented, but which are not ready yet.

1.4 Natural-language question-answering system

The natural-language question-answering system (QuickAsk) [8], [3], [4] used in the web site is based on templates. For each answer, one or more templates are developed, which will match many different variations of questions, for which this answer is suitable.

Example of a template:

$eat $food ; sensibl* rational* levelhead* reasonab* unreasonb* prudent* intelli* sane* insane* unrealist* realist* thoughtful* credib* understand* know* clearhead* bright* perspica* precept* astut* smart* apt suitab* witt* shrewd* $good together party* partie* band* company* bunch* group* gathering* alone lone* solitar* gregarious* secluded* single* desolate* separat* friend* accompan* unaccompan* [on ; $people ; own] [with by at in # $people]

This template will match for example the following questions:

- What eating is sensible?
- Help me with problems with eating at parties.
- Suggest a smart diet.

This system requires that one or more such templates is constructed for each answer. During usage, the questions asked by actual users are logged, and these log files are used to check if the system produces suitable answers – when not, either the templates may need to be revised or a new informational text written. The total time spent on producing a good template for each answer is 15-60 minutes, including time spent on testing the templates and on revising them based on usage logs.

An alternative would be to use traditional so-called free-text search tools, which automatically match questions to words in the answers. The advantage with the system we used is that it more often will find the best answer to a question, and that the search response will contain less unsuitable answers (in information retrieval terminology, our system will give higher recall and precision than free-text search tools). The advantage with free-text search tools, of course, is that the manual work of producing the templates is not needed.

Two master's students at DSV [12] have compared this system with using Google with site-restriction to only “site:web4health.info/sv/”. The test was done with 50 randomly selected actual questions from the log files of questions asked to the system. They found that the natural-language question-answering system found a good answer to 90 % of all questions, compared to only 68 % for Google. Traditional so-called free text search systems were usually less good than Google, except for SiteSeeker, which was slightly better than Google (72 %) but still not at all as good as the natural-language question-answering system QuickAsk. SiteSeeker achieves better results than Google by understanding misspellings and Swedish-language conjugations better than Google.

2 The KOM2002 content development system

The KOM2002 content development system has the following functionalities:

Handling of objects, which can be split into fields. Each object can exist in more than one language. Fields can be marked as being identical in all languages, having possibly different text in different languages, and in the latter case whether Systran machine translation [7] provides an initial text as an aid to the human translator producing a new translation (all pages are translated by humans, but machine translation can be used as input to the human translator), whether a field is mandatory, etc.

Objects can be linked to each other, for example all pages are linked to an area where page texts are stored, other texts, like synonym lists and stop lists for the natural-language answering engine, are linked to an area for export objects other than texts. Comments and discussion of a page, internally between the developers, can be stored in a separate discussion area associated with each page being developed. More complex link structures can be defined for handling of, for example, work flow applications.

When an object is to be translated, the translator locates the source version, and chooses the target language. A window is then opened, which shows the text in both languages side by side. Some of the fields are automatically filled with Systran translations (See Figure 1). The translator can then supply the text in the target language, and, when ready, submit the translation.

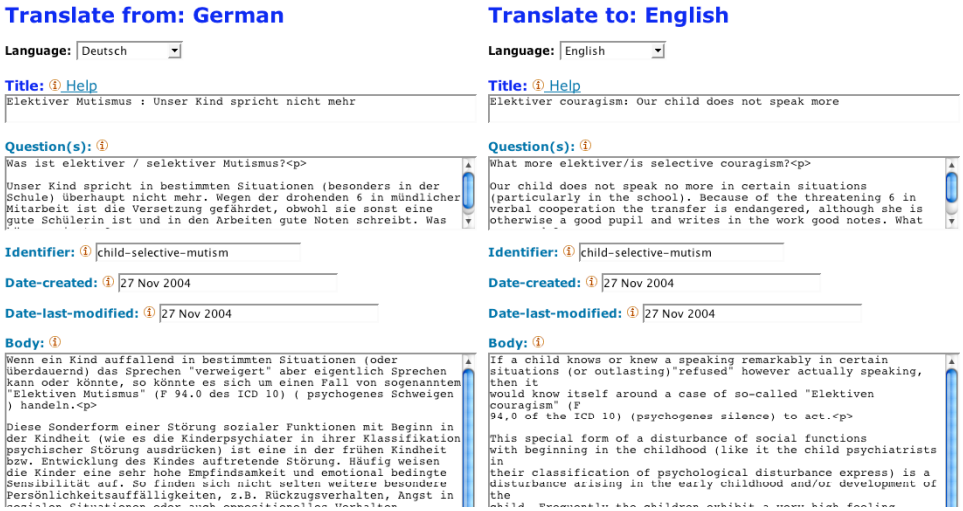

Figure 1: Part of the window when doing a translation

The translation window can also be used when modifying a translation. For example, if a change has been made in the German original, and an English translation already exists, the translation window can be used to see the new German text side by side with a window where the change can be copied to the English translation. This feature is especially valuable for changes in the natural-language classification, since this is often changed in order to better cater to experience from the system logs of how well the natural-language question-answering system works.

The content-management system also has commands to export pages, when ready, to the static pages available to external visitors, and to information in the data base used by the natural-language question-answering system. The same page is usually exported to multiple exported pages, such as a page for screen viewing, a page for printing, a page with a list of sources and a page as stored in the data base of the natural-language question-answering system.

The content management system also has a forum and chat facility which can be used by both developers and external visitors.

The system has a compare facility, which shows the changes between two versions of the same page, and a facility to checkmark an object, while one of the editors is working on it, to prevent more than one editor from modifying the same object at the same time.

3 Multi-lingual information

Examples of information which needs to be translated to multiple languages:

A. Attributes of informational pages:

1. Titles.
2. Questions.
3. Answers.
4. Source references.
5. Author name.

6. Meta-description (for search engines, some of which want a short summary of each page).
7. Meta-keywords (used by some search engines).
8. Classification for the natural-language question-answering system.

B. Other texts:

1. Synonym lists used by the natural-language question-answering system.
2. Stop lists used by the natural-language question-answering system.
3. Export templates. The data base information for each page is inserted in such a template to produce the page shown to external users.
4. User interface pages and phrases, including the home page for each language.

4 Experience with multi-lingual content-management

Each web site has its own data structures of objects linked in different ways. At the leaf end of these data structures are often texts which have to be translated for a multi-lingual web site. Some editing operations will only change the structure, and not the text leaves. A good multi-lingual system should allow editors to make such operations only once, and have immediate effect in all languages, by separating language-dependent texts from language-independent structure.

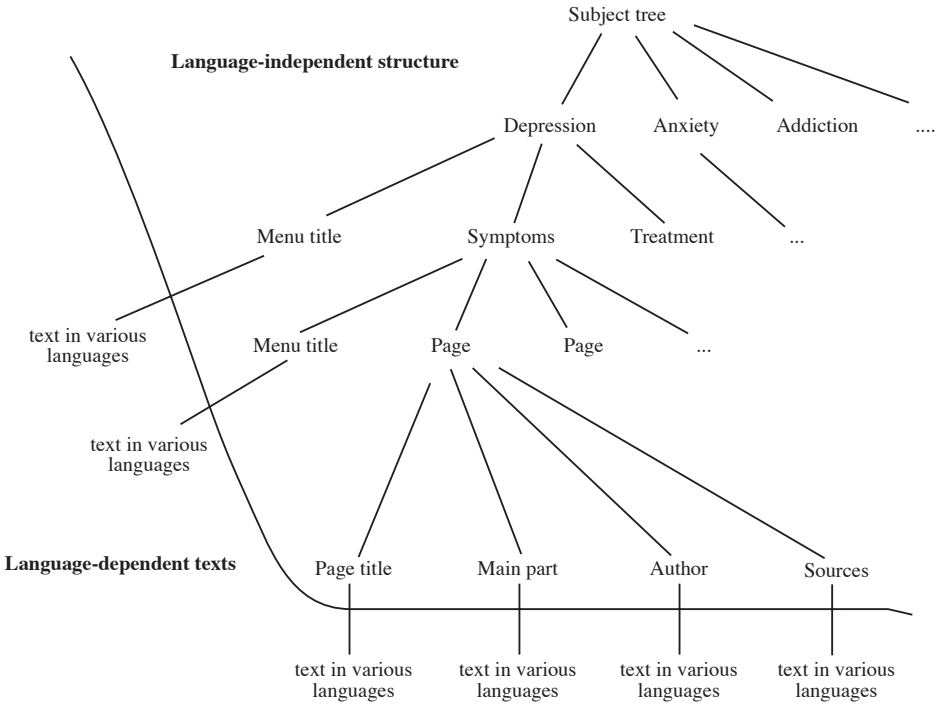

Figure 2: Separation of language-independent structure from language-dependent texts.

Figure 2 shows an example of a hierarchical structure, where the structure is language-independent and only the names in the leaves need to be translated. But the same principle applies to all kinds of structures. If, for example, a web page contains a section with links to related pages within the web site, then only the user-visible strings need be translated, the structual linking-information need not be translated.

This is especially important when step-wise improvements are done to the content of a web site. A change of the structure should then automatically be available in all languages, not only in the language in which the change is made. This is very important in order to retain high quality in a site where step-wise improvement is done.

Other editing operations need creation of one or more new texts. A good multi-lingual system should allow such operations with only one single operation to create the object in one initial language, plus added operations to translate the new texts to each target language. Our experts usually either write their original texts in their native language, and then translate it to English, or write the original texts directly in English. Other experts can then translate it, usually from English to their native language.

When visitors view the web site in their native language, the system can either be designed so that they will only see texts available in their native language, or so that texts which have not yet been translated are shown in another language, usually English. We have chosen the second option - whenever a text is not available in the native language, the English version is shown instead. A third alternative might be to let Systran or some other machine translation engine translate the texts from English, possibly as a temporary measure until a human has made a better translation. We have chosen not to do this, since some people are offended by machine translations. This seems to be a personal thing, some people think they are quite useful, even if not always perfect translations, others cannot accept imperfect language at all.

Often, the task of creation of structure is done by other people than those who translate the texts to different languages. Some work flow functionality is then useful. The most important work flow functionality is a tool, by which a translator can find which new texts need translation.

We defined our system with the goal of having the choice of language at the outermost structure of the data structures, as described above. The system mostly adheres to this principle. However, the system does not yet have built-in support for language-independent handling of the hierarchical subject trees (taxonomies). This means that creation and translation of such structures is not at present as easy as it should be.

As much information as possible should be specified in only one language. Thus names of objects which are only visible to the developers are always in English. Only the texts visible to users need be available in multiple languages. This makes translation easier. For example, the synonym lists used by the natural-language question-answering system have all the names for the synonyms in English, only the values need vary between languages (see Table 1).

Synonym name	Value in English	Value in German	Values in other languages
$adhd	[ad ; hd*] adhd* ahdh* adhs* [a ; d ; h ; d] [d ; a ; m ; p] [a ; d ; d] [attention;deficit;hyperactiv*;disord*] hyperactiv* hypercinet* addh* damp* adhs* twitch*	[aufmerksamkeits ; defizit ; syndrom] [ad ; hs] [a ; d ; h ; s] ads adhs adhd hks hyperaktivitaet* zappel* hampeln hyperkin* konzentrationsschw* unaufmerksam* ablenkbarkeit* impulsivit* verhaltensstö* entwicklungsstör* unkonzen* abgelen* tagträum* träumer* chaosprinz* zerstreut*	...
$advantage	advantag* pro pros prefer* benefi* asset* gain* favor* favour* positiv* good* triumph* succe* excel*	bevorzugung dienlich*, einträglich*, ergiebig* ertrag frucht gewinn interes* nutz* oberhand oberwasser plus, vorzu* vorteil* guenstig* posit* gut* besser erfolg*	...
$anorexia	ana anore* anero* anere* anerx* anorx* starv* [no not ;hung*] undernourish* fast fasting fasted tiny little petite weigh* apath* slinky lean meager gaunt lanky skinny famine famish* drought unfed meager* thin gracil* svelt* willow* thin slender* [low : weight] slender* [not ; want* wish* desir* like* : to ; $eat] [refus* declin* reject* rebuf* : to ; $eat] underweig*	anorexie* magersucht* magersuecht* magersuecht* duerr* kachex* kachekt* hager schmal ausgehunger* ausgemergel* unterernaehr* untergewi* abgezehrt spindelduerr abgemage* abgez* arid dünn dürr gertenschlank hager knochendürr knochig, kümmerlich rappeldürr schlank, schlankwüchsig schmächtig spindeldürr hohlwangig	...

Table 1: Part of the synonym list. Note that the names of the synonyms, visible only to the developers, are in English for all the languages.

5 Cross-lingual natural-language question-answering

As described above, the natural-language question-answering method we have used means that we have to produce question-matching templates for each page. These templates also often need to be updated, based on entries in the usage logs where the system did not provide the best answer to a certain question. The work of developing and managing these templates require a special competence. Not even an ordinary professional translator can do it without a few days of instruction on how to create such templates.

Because of this, it is an advantage if only some of the people need to have this particular competence. Also, it is very important that a change in these templates can be done in one language, and the result be immediately available for natural-language question-answering also in other languages.

We have implemented this, using a technique called cross-lingual natural-language question-answering [1], [5]. How this works is shown in Figure 3. The figure uses

Italian, but Italian can be replaced by any other language, for which a machine-translator to English is available.

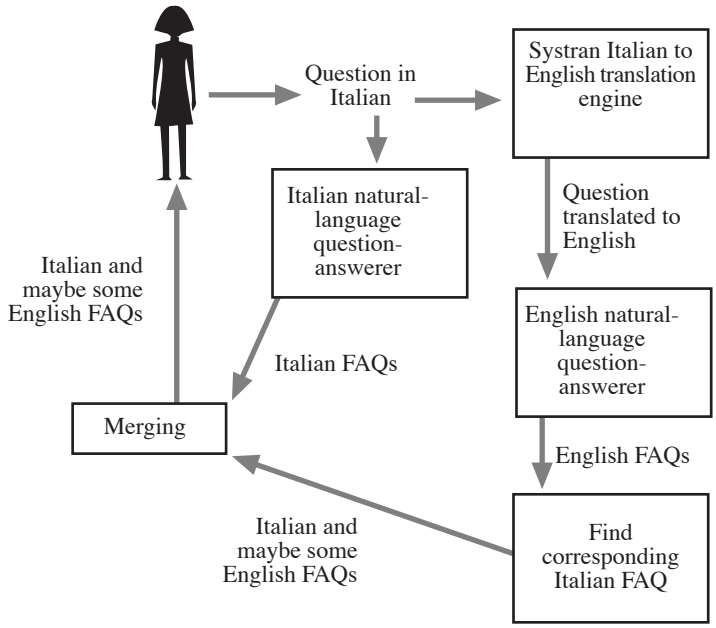

Figure 3: Cross-Lingual Natural-Language Question-Answering.

Incoming questions are translated by Systran machine-translation to English. The English question is then put to the English-language answering engine. When the results have been found, the corresponding native language objects are shown. This could be implemented so that the user never sees that any other language than his own is involved. We have chosen, however, to show the English answer if the text of the answer has not yet been translated to English. This means that users will see some English answers mixed with their native language answers.

It is also possible to set up this process without having any translated answers. This will allow users to ask questions in their native language, but get the answer in English. Since many people handle English better as a passive than as an active language, this would be a useful tool for them.

We also have some texts which are only available in the native language, since each national editor can add texts which are only available in his/her own language. For these texts, a native-language question-answering system is used to find answers.

We have compared the quality of the answers found in this way to question-answering directly in the language of the questions [6]. These comparisons indicate that taking the Systran machine-translation engine as is, the quality will be somewhat inferior to that of direct language answering. However, if the dictionary used by Systran is extended with the terminology suitable for our subject area, the quality will be almost as good as with direct language answering.

The reason for this is that the standard Systran dictionaries are designed for office documents, not health. For example, the word “body” is by Systran translated as if it

meant “main part”, which is the most common use of this word in office documents, but which, of course, is usually not suitable when talking about health.

One might argue that augmenting the dictionary with new terminology (in our case about 6000 words, but many of them already have suitable translation in Systran, so all need not be added in an additional dictionary) is as much work as writing the classification separately in each language. However, this is not true, because the same dictionary entry can be used in the classification of many answers. For example, the dictionary entry for “cause” can be used in many pages discussing causes of various disorders. Another important advantage with cross-lingual question-answering is that development of the dictionary does not need the special competence needed for doing the classification. Thus, cross-lingual question-answering allows a separation of tasks between people with different competences.

6 Work Flow and News Control

In a multi-lingual web site, different people perform different tasks. In our case, medical experts write the texts. Each text is also checked by another medical expert than the original author. They are translated either by other medical experts, or by other translators, where the translation is checked by a medical expert. The classification and most of the structuring is done by linguistic or computer-science experts. The same text often has to pass through many hands before being finally published in each target language. It is then important to have so-called work flow support.

Our system has a work-flow system, which can be configured for different work flows. The most complex work-flow presently supported has the following states:

1. The Swedish medical expert writes a Swedish answer to an English question.
2. A translator translates the Swedish answer to English.
3. The Swedish medical experts checks the translation for correctness.
4. The translation is checked by an English-language medical expert.
5. The translation is published.

The work flow system should it easy to see the state of an answer, and to get a list of all answers which are in a special state. It should also give content developers easy feedback on when some work should be done, and reminders when this has not been done.

The content management system also has tools to aid the transformation of an answer to a personal user question, written by a medical expert, into a general answering page which other users can find when searching the data base. This transformation involves creating templates for the natural-language question-answering system and other similar activities. Long questions are often abbreviated at this stage. The system marks which such user questions have been converted, and which informational page is a conversion of which answer to a specific user question.

Work flow notes the stage of each piece of information, what further action is needed, and notifies the approporiate person who is to perform a certain act. Knowing what each person is expected to do, and reminding them of tasks left to do, is known under the term ”news control” and has some similarities to the capabilities of mail programs of knowing which emails a person has not yet ready.

7 Left to do

Here is a short list of functions which we are not ready with, but which we think would be an improvement to our tool:

1. A WYSIWYG (What You See Is What You Get) editor like Dreamweaver for editing texts and translations. Since the content developers use a web-based interface, this tool should be an applet, so that users need not install additional software on their computers. We have developed two such editors, one written in Javascript, the other written in Java, but have not yet got them working in our production environment.
2. More complete support to the principle of separating structure and text is needed.
3. A spell checker, so that also badly spelled questions can be answered. Again, we have this developed, but not yet entered into our production environment.
4. A tool to ease the migration of extensions of an already existing answers to other languages. This tool should show to the translator exactly what has been changed in the source language version of an answer, when translating this change to the translation. We already have the tool to show differences between versions, but have not yet implemented it in a neat way for migration of changes from language to language.
5. We have found that for workflows with long chains of steps taken by different people, there is a risk that objects get stuck. Thus, it would be useful for a tool to find or remind system administrator when an object is stuck in a step of an unfinished workflow, and to remind content developers when some task is expected of them. We have such a facility, but it needs improvements.
6. The multi-lingual question-answering system would work better, if Systran was given a special dictionary of terms used in question templates.

Just now, we do not have funding for doing these improvements, but we hope to get it in the future.

8 Conclusions

The main conclusion of this development is that it is important to design multi-lingual systems so that language-independent structure is clearly separated from language-dependent texts, so that changes in the language-independent structure can be done for all languages in one operation. This is achieved by putting the texts at the leaves of the data structures, and designing the system so that each such text-leaf can easily be specified in multiple languages and easily be translated. Important is also support for the work flow between different people doing different tasks, such as writing texts and doing the translation to different target languages.

9 References

[1] Cross-Language Evaluation Forum - CLEF, by Michael Kluck, http://www.gesis.org/en/research/information_technology/CLEF_DELOS.htm

[2] Professional Content Management Systems: Handling Digital Media Assets, by Andreas Mauthe and Peter Thomas, ISBN: 0-470-85542-8, Wiley March 2004.

[3] Web4Health Complete Final Project Report, by Jacob Palme, July 2004, http://web4health.info/documentation/D-7-4-full-final-rep.pdf

[4] Natural Language Question Answering System Classification Manual by Jacob

Palme and Eriks Sneiders, http://web4health.info/documentation/D2-2b-classification.pdf

[5] CLEF - Cross-Language Evaluation Forum, by Carol Peters, http://www.ercim.org/publication/Ercim_News/enw40/peters.html

[6] Sjödin, Elin: Preliminary Results (extracts from a forthcoming Master's thesis, Stockholm University).

[7] Access to the online translation engine, by Elsa Sklavounou, KOM2002 project report D 8.1 December 2002, http://web4health.info/documentation/D-8-1-translator-access.pdf

[8] Automated FAQ Answering: Continued Experience with Shallow Language Understanding. Question Answering Systems by Erik Sneiders. Papers from the 1999 AAAI Fall Symposium. Technical Report FS-99-02, November 5-7, North Falmouth, Massachusetts, USA, AAAI Press, pp.97-107 at http://www.dsv.su.se/~eriks/Sneiders1999.pdf

[9] Short sample of: The CMS-Report Web Content Management Products & Practices", by Tony Byrne, CMS Watch (www.cmswatch.com), autumn 2002.

[10] "Content Management Systems: Getting from Concept to Reality", by C. Kartchner, The Journal of Electronic Publishing June 1998, Volume 3, Issue 4

[11] Content Management Bible", by Boiko Bob, John Wiley & Sons; 1st edition, December 2001)

[12] Utvärdering av hälsosajter (in Swedish, title translated to English: "Evaluation of health web sites"), by Ideh Alikhani & Bushra Al Hamdani, Master's thesis at DSV June 2005.