Linking a Norwegian web portal for Language Technology to its Nordic partner sites

by

Koenraad de Smedt and Gisle Andersen
AKSIS / University of Bergen

1. Introduction

In order to achieve a truly Nordic perspective, the Norwegian Documentation Centre for Language Technology has during its four years of activities (2001-2004) intensively cooperated with the other Nordic documentation centres in Denmark, Finland, Iceland and Sweden. This cooperation has eventually crystallized into a system for linking together the various national web portals. Located at http://www.norskdok.uib.no/, the Norwegian portal is aimed at providing a news service as well as a comprehensive and updated survey of activities in the field, language resources, networks and contact information of the participants. This information has been made searchable across all the Nordic documentation centres.

2. Organization and meetings in 2004

Organisationally, this year saw a change of staff at the Norwegian Documentation Centre, as Gisle Andersen replaced Kristin Bech as Project Manager and responsible for the everyday documentation activities and practical matters at the centre. The Norwegian team comprised of Gisle Andersen as project manager, Koenraad de Smedt as scientific coordinator, and Torbjørg Breivik representing a link to Norsk Språkråd (the Norwegian Language

Council). This team participated at two Nordic network meetings, in Copenhagen on April 23, and in Helsinki on October 8, and a liaison meeting with LT World in Saarbrücken, Germany, on December 13, where it was agreed to cooperate closely in activities concerning language technology terminology. Furthermore, Koenraad de Smedt represented the Norwegian Documentation Centre on the Nordic documentation centres' joint visit to five universities in the Baltic countries (Vilnius, Kaunas, Riga, Tartu and Tallinn) in October.

3. The Norwegian web portal

In order to effectively document current activities and results, a substantial amount of content has been added to the Norwegian language technology portal, informally known as NorskDok. Figure 1 shows the front page of the portal. Its main text frame on the right of the page displays a welcome message and introduces the Norwegian Documentation Centre for Language Technology. At the top is a main menu showing the main content categories. At the left, there is a local menu, a search window and an overview of the five latest news items.

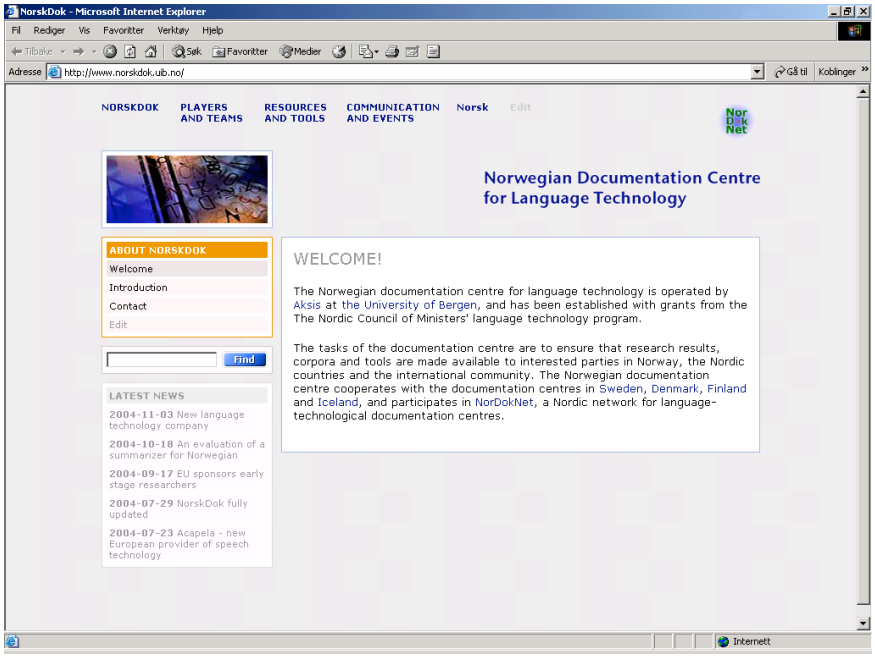

Figure 1: The Norwegian Documentation Centre for Language Technology

Although the visual layout is local to the Norwegian site, the structure of the information, as reflect in the various menus and submenus, are based on the common Nordic categories that have been agreed upon in earlier NorDokNet network meetings: *people, organizations, companies, projects, materials, products, research systems* and *news*.

The actual information retrieval took place in several steps; first the web and other sources were consulted, especially the web pages of universities and other R&D institutions, the Norwegian Research Council, and private companies, but also printed sources such as *Språkbankrapporten* (Report on a Norwegian Language Bank) and, indeed, the earlier volumes in the current series of yearbooks. This yielded a lot of information concerning ongoing and former projects and resources, from which the most important information could be extracted. A small amount of information was added in response to email sent out to a comprehensive list of contact persons.

The database accessible through the web portal is implemented as a MySQL database, in which XML is stored and retrieved dynamically. A uniform graphical layout of the web pages was ensured by means of XSL transformations into HTML and by CSS style sheets. The information for individual items in the database can be produced and edited by authorized administrative users through user-friendly web forms, exemplified in Figure 2.

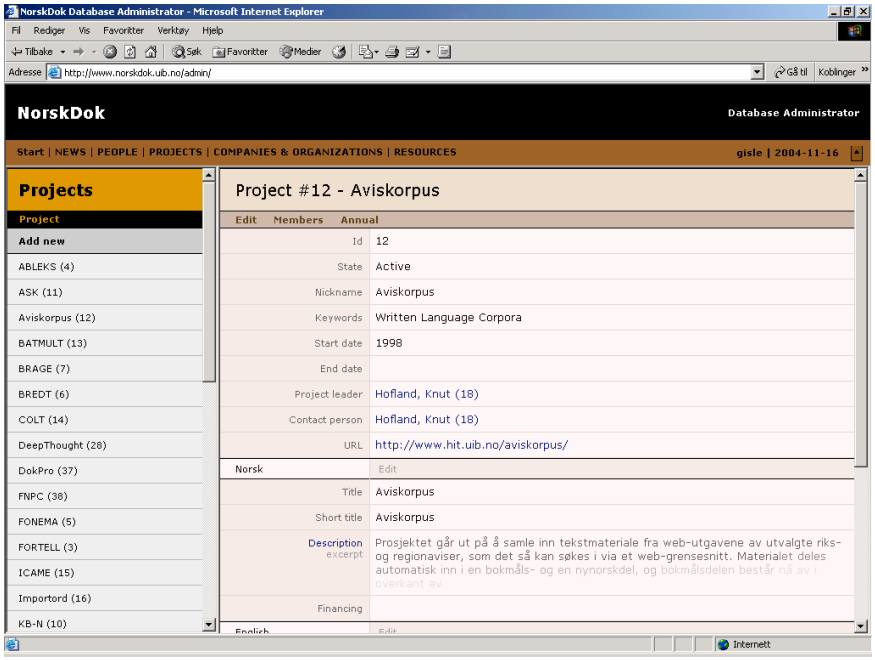

Figure 2: A database entry form

During 2004, many new database entries have been added, especially involving companies and products that were hitherto not represented. The portal is bilingual, so each page in Norwegian has an English counterpart. The overall result of this work is an updated web portal which we believe to be the most comprehensive existing database providing current information on Norwegian language technology resources and participants.

The task of keeping pages for bread text updated is simplified by the fact that a link on the web page gives administrative users access to a window for editing the XML, as shown in Figure 3.

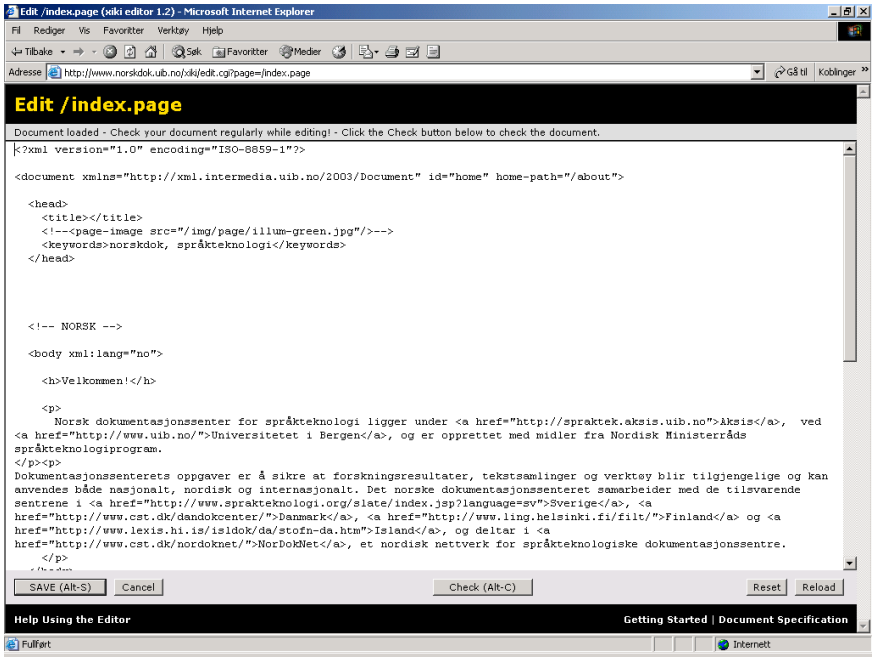

Figure 3: The edit window

The edit window encompasses an easy-to-use XML tagset, defined in the DTD, specifying headers, lists, emphasis markers, etc. The tagset, menu items, DTD etc. are all defined locally, which gives the user full control of the various components of the database. It is also possible to upload single files or several files at the same time (using a Xiki file exchange tool), and to edit text files on a local machine, e.g. by means of a Perl script or otherwise, thus reducing the need for manual work on individual pages.

4. Linking to a Nordic search tool

In order to achieve a Nordic perspective of the documentation effort, a search tool was developed jointly by all the Nordic documentation centres, with the aid of Hercules Dalianis at SiteSeeker. The requirements for this search, which were agreed upon in 2003, can be summarized as follows:

1. allow search of arbitrary terms in full text and of predefined keywords (based on the LT World keyword list) in *meta* tags.

2. optionally restrict the search to one or more given categories: people, companies, organizations, projects, materials, products, research systems, news.

In order to satisfy the latter requirement, the pages related to the different categories were stored in separate directories. In order to achieve the former requirement, each individual page was annotated with one or several of the agreed keywords describing the subfield to which the activity, person or project is connected (such as *speech synthesis, machine translation,* etc.). The keywords are coded in the *meta* tags, so that a web page may yield a hit even though the search word is not mentioned in the visible text of that page. On the other hand, search words that are not predefined keywords can still be searched in the full text related to the database item.

The search across all the Nordic sites was implemented by means of the search tool *SiteSeeker,* which allows searches over a number of different sites simultaneously, and allows the definition of categories. For each information category, e.g. *companies,* SiteSeeker was instructed at which location it could find the pages for Danish companies, at which location the pages for Finnish companies, at which location those for Norwegian companies, etc. The indexing of all pages is automatically updated each night. Provided that the keywords in the *meta* tags are coded consistently, it is therefore possible to answer to queries such as e.g. (1)-(3). Search (1) is exemplified in Figure 4. Note that only keywords in English have been added, and that it is not yet possible to use non-English keywords to search across different languages.

(1) Which Nordic projects or persons deal with machine translation?

(2) What is the Nordic news on speech synthesis?

(3) Which written language corpora are available in the Nordic area?

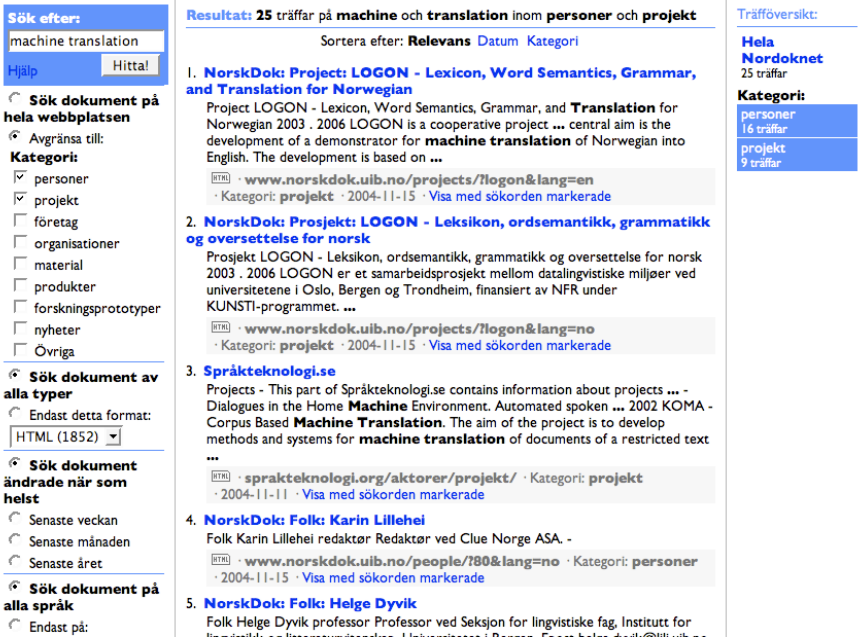

Figure 4: Search Nordic projects or persons dealing with machine translation

The web portal has been announced nationally, and we have established an e-mail list of contacts, generated automatically on the basis of contact details in the database. The list is used as a means of getting in contact with different players in the field, and for requesting updated information about ongoing activities, news etc. We are pleased to have received generally positive responses, and some pages could be updated on the basis of highly useful feedback from list members.

5. Planned activities in 2005

Now that the current project is nearing completion, our concern is to secure the material at the Norwegian Documentation Centre for future use and to continue and even broaden the documentation effort in cooperation with others. We consider it a national obligation to keep the language technology database as updated as needed for its effective use. A continued Norwegian language technology portal will constitute a useful point of departure in the possible establishment of a future Norwegian Language Bank, for which plans exist. We are also currently considering the inclusion of information for and about

students, such as ongoing student projects and information for and about job seekers within the field.

Our October contacts with language technology research groups in Lithuania, Latvia and Estonia have revealed a need for documentation of Language Technology in these countries, and a willingness to carry out such efforts in cooperation with the current Nordic documentation centres. A broadening of the international cooperation effort, especially in the Nordic-Baltic region, is a highly desirable goal.

We will also seek further cooperation with LT World (http://www.lt-world.org), especially with regard to implementing cross-lingual searches in several languages. In this context, we are planning further activities to ensure uniformity and clarity in the use of terminology for language technology. In this connection we are hoping to continue our good cooperation with the Norwegian Language Council, the other Nordic documentation centres, and LT World. Our aim is to compile and disseminate a multilingual, quality assured terminology list and to ensure that the terms acquire an official status and are used by Norwegian language technology participants in teaching as well as research.