Publishing multilingual ontologies: a quick way of obtaining feedback

J. JOÃO ALMEIDA; ALBERTO SIMÕES

Departamento de Informática
Universidade do Minho
Campus de Gualtar
4710-057 Braga, PORTUGAL
e-mails: jj@di.uminho.pt, ambs@di.uminho.pt

Keywords: ontology publishing, thesauri publishing, dictionary publishing

Dictionary and Thesaurus are valuable resources for Natural Language Processing but do not exist as freely available as expected, especially for languages other than English and, when they exist, they are just available for querying online.

Our main goal with T2O -- Thesaurus to Ontology framework --- is to create a multilingual ontology:

- freely available online and to download;
- with a computer readable format;
- with a good API;
- with a structure as rich as possible;
- reusing all the structured information we can get;

The main approach to get into these objectives is to define an ontology algebra to deal with the main object type (ontologies) but as well other objects we will want to reuse (thesauri, taxonomies and word lists). This ontology algebra includes these major operators:

- translation: add a new language to a thesaurus which does not include it;
- inversion: take a thesaurus which has a specific base language and other languages and change its base language to one of the available languages;
- completion: definition of a set of rules to be able to auto-complete the ontology, like inversion rules (specify that two relations are inverses one of the other) and forward-chaining rules.
- assimilation and joining thesaurus: take different sources thesaurus and join term properties, as well as treat ambiguities and different concepts identified by the same term.
- publishing: produce views or ways to browse and search the ontology created.

This poster main focus is on the last operation. Just to prepare a resource is not sufficient. The most important part in these kind of projects is to make the resources available in a suitable format, useful for the main user.

Is it not just important to publish, but also to do that early. To publish shows the utility of things earlier. If you need to increase the budget of your project, for instance, it will help if you can show how things will look at the end of the project. Also, people start looking at the published documents, giving feedback. This feedback will allow you to develop better resources, best views as well as to improve usability of the views made available. Finally, to publish help you to get friends and collaboration for you projects. Lot of projects began with little information and now include gigabytes or data. If you wait to have a good quantity of information before
publishing it you will end up with a slower grow curve.

In this project we developed four main publishing formats:

PAPER
We are in the electronic era, but we all know that paper is still used and will continue to be. To publish in paper we need to produce a printable document and, as portability matters, the format chosen was PDF. The process of publishing in paper produces a LaTeX file which, after compilation, produces a PDF file ready to be printed.

These files graphical format is very similar to a standard paper dictionary, which means this tool can be used to produce high quality dictionaries for real publishing.

This publishing format is interesting in case you are working for a publisher and needs really to print a dictionary, or if you are working with small ontologies. For our project ontology, with more than fifty thousand terms, we would need more than a thousand sheets of paper.

WEB
Exporting the ontology to HTML and making it available for browsing in the Internet is crucial. To perform this task we used two approaches:

1. serve dynamic pages statistically: automatically generate an HTML file for each term in each language from the ontology, and create hyperlinks between them. This approach has the advantage of high-availability, but the disadvantage of the number of HTML files created, as well as the space needed for those files.
2. really serve the pages dynamically: create a database with the ontology terms, and HTML parts of the main term pages, and then construct pages on-the-fly. This approach has the advantage of easy search facilities.

So, it is important not only to make these pages available but as well index them and let the user navigate and query directly the database.

WIKI
The idea behind this format is not properly to publish the ontology, but to enrich it. Everybody knows the success of wikipedia and related projects. If we can have a wiki to let people access the ontology terms, and edit the relations, it can grow easily.

Also, as we developed some Domain Specific Languages for quick thesaurus creation, a simple edition of a wiki page can create some dozens of terms. We are using a simple tabular format to define word lists and their characteristics, and then transform the tables into ontology entries.

DICTIONARY SOFTWARE
From the old dictd dictionary server, to the more recent stardict, wiseDict and xdxf open-source dictionary projects, our ontology format can be easily transformed in any of the file types used by these software tools.

The ability to produce resources for these tools is important, given that this way users can just download the dictionary and consult them offline.