Dictionary Building with the Jibiki Platform: the GDEF case

Mathieu Mangeot*, Antoine Chalvin†

* LISTIC
F-73376 Le Bourget du Lac Cedex
Mathieu.Mangeot@univ-savoie.fr
†INALCO
2 rue de Lille F-75343 PARIS CEDEX 07
Antoine.Chalvin@inalco.fr

Abstract

This paper presents the use of the "Jibiki" generic dictionary online development platform in the case of the GDEF Estonian-French bilingual dictionary building project. This platform has been developed mainly by Mathieu Mangeot and Gilles Sérasset based on their research work in the domain. The platform is generic and thus can be used in (almost) any kind of dictionary development project from simple monolingual lexicons to complex multilingual pivot dictionaries as well as terminological resources. The platform is available online, thus it allows entry writers to work and collaborate from any part of the world. It consists in two main modules and data management tools. There is one module for elaborating complex queries on the data and one module for editing entries online. The editing modules generates automatically an interface from the XML structure of the entry.

1. Introduction

In this paper, we presents the "Jibiki" generic dictionary online development platform used for the GDEF Estonian-French bilingual dictionary building project. This platform has been developped mainly by Mathieu Mangeot and Gilles Sérasset based on their research work in the domain. The platform is generic and thus can be used in (almost) any kind of dictionary development project from simple monolingual lexicons to complex multilingual pivot dictionaries as well as terminological resources. The genericity in the dictionary lookup is obtained thanks to the definition of search criteria based on Xpath strings in the entries structure. For the entry editing, it is obtained thanks to the automatic generation of the editing interface from the XML entry structure. The platform is available online, it allows entry writers to work and collaborate from any part of the world.

We begin by presenting an overview of the platform. Then we list 3 projects that are currently using the platform. It consists in two main modules and data management tools. There is one module for elaborating complex queries on the data and one module for editing entries online. The next sections detail the platform usage: first, the dictionary lookup module, then the entry writing module and finally, some task management tools.

2. Overview of the platform

The Jibiki platform (Sérasset, 2004) is an online[1] generic environment for writing and querying all kinds of dictionaries: terminological glossaries, bilingual dictionaries, multilingual lexical databases, etc.

It has been developed mainly by Mathieu Mangeot (Université de Savoie, France) and Gilles Sérasset (Université de Grenoble 1, France), thanks to research driven by the GETA team of the CLIPS laboratory in Grenoble, France (Sérasset, 1994; Mangeot, 2001).

The platform is implemented in Java, exclusively with open source tools. It is based on Enhydra, a web server of dynamic java objects and Postgres, a relational database. The interface is available in English, Estonian, French, German and Japanese. New languages can be easily added. The snapshots of the Figures 2 and 3 were taken with the English interface and the ones of Figures 3 and 4 with the French one.

Annex tools have been added on various instances of the platform. Some facilitate the communication between communities of users (forums, distribution lists) and others, the work of the lexicographers (tool for managing aligned bilingual corpora).

3. Projects currently using the platform

The platform is currently used by three lexicographical or terminological projects.

3.1. Papillon Project

This project[2] (Mangeot et al., 2004), launched in 2001, is at the origin of the building of the platform. Its main goal is the construction of a multilingual lexical database with a pivot structure covering among others the following languages: Chinese, English, French, German, Japanese, Lao, Malay, Thai and Vietnamese. The resulting resources are publicly available and free of rights. The projects is open to all those who are interested in these languages.

3.2. GDEF Project

The GDEF (Great Estonian-French Dictionary) project[3] (Chalvin and Mangeot, 2006) started in 2003. Its goal is to build a bilingual Estonian-French dictionary of about 80,000 entries, by a team of 8 people made of linguists, as well as Estonian and French translators.

The reflections conducted in the GDEF project oriented in a decisive manner, in 2004 and 2005, the development of

[1] http://jibiki.univ-savoie.fr/jibiki

[2] http://www.papillon-dictionary.org

[3] http://www.estfra.ee

the platform. A more precise definition of the needs resulting of a professional lexicographic work lead to numerous new functionalities. The tests driven by the GDEF lexicographers strongly contributed to transform an experimental research prototype into a fully functional platform.
The GDEF is now the most active project using the Papillon platform. It is supported by the AIF (Intergovernmental Agency for Francophony), Robert Schuman Foundation and the embassy of France in Estonia.

3.3. LexALP Project

The European project LexALP[4] (Sérasset, 2005), was launched in 2005. Its goal is to harmonize the terminology of the Alpine Convention four languages (French, German, Italian and Slovenian) so that member states are able to cooperate effectively. For this, the project uses the Jibiki platform in order to build a term bank used to compare the specialized terminology of six different national legal systems in four different language, and to harmonize it, optimizing the understanding between various alpine states in environmental matters at a supranational level.

4. Dictionary lookup

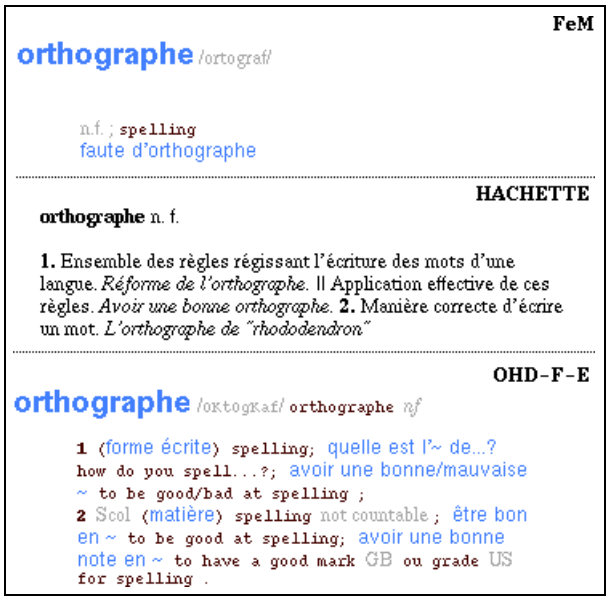


Figure 1: Result of the query of the French word "orthographe" in multiple dictionaries

The platform allows one to lookup all the dictionaries available on the server and to display the results in the same window like in Figure 1. The advanced query interface shown in Figure 2 offers a combination of multiple search criteria on:

- the languages: source, targets, available resources;
- the character string: prefix, suffix, substring;
- the content of the entries: headword, variants, pronunciation, domain, gloss, part-of-speech, translations, examples, etc.

When elaborating a query, the user can interactively add a search criteria by clicking on the "+" button. If the available criteria are not sufficient, it is even possible to define new search criteria when a new resources is added by defining common pointers on searchable information parts. These pointers are defined with XPath strings on the XML entry structure.
In the case where a normal search returned no results, a reverse lookup is also executed. For example, when a user looks for the Estonian translation of a French word but it is not available into French->Estonian resources, the system will also lookup into Estonian->French resources and the Estonian entries containing the French word will be displayed.
The users can also define their own style sheet for viewing the search results.

5. Entries Writing

The writing of the entries is done directly online on the platform via a web browser. The writing interface is generated automatically from the description of the structure of the entries (an XML schema), thus allowing the edition of (almost) any type of dictionary entry as long as it is encoded in XML. A preliminary version of this module has been developed by (Mangeot and Thevenin, 2004) from previous research done on plasticity of user interfaces. The actual version of the module was then simplified and rewritten from scratch.
The interface is shown in Figure 3. It is built upon an HTML form and can also deal with relatively complex structures thanks to more elaborated interactors that combine the basic HTML ones (text boxes, radio buttons, popup menus). Such example is the list management one that allows the writers to add, delete or reorder elements in a list by simply clicking on a button. These elements can be themselves complex objects containing lists of other objects, etc.
A specific module allows the writer to establish links to entries in other resources available on the server. This technique is mainly used for linking an entry to its translation in another language when the translation already exist as a separate entry or for building a pivot acception that links entries in different languages.
Every change made in the entry is stored in a history. It is then possible to come back to any previous version of the entry just like the usual "undo" commands.
In the GDEF, it is used in order to display some information concerning the French equivalents (gender, irregular forms, etc.).
The writing process is divided in several steps depending on the project. The GDEF is the most complete with three steps:

1. A contributor writes an entry;
2. It is next revised by a reviewer;

[4]LexALP: Legal Language Harmonisation System for Environment and Spatial Planning within the Multilingual Alps

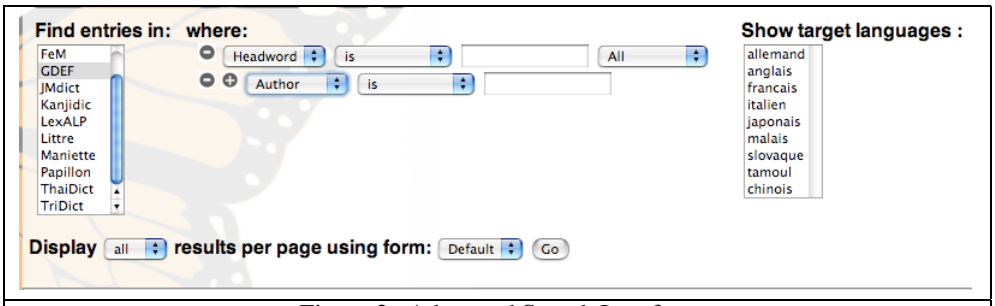


Figure 2: Advanced Search Interface

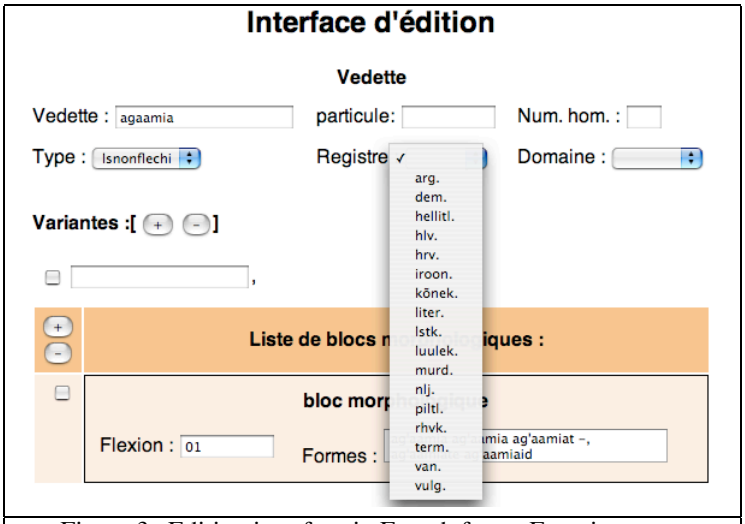


Figure 3: Editing interface in French for an Estonian entry

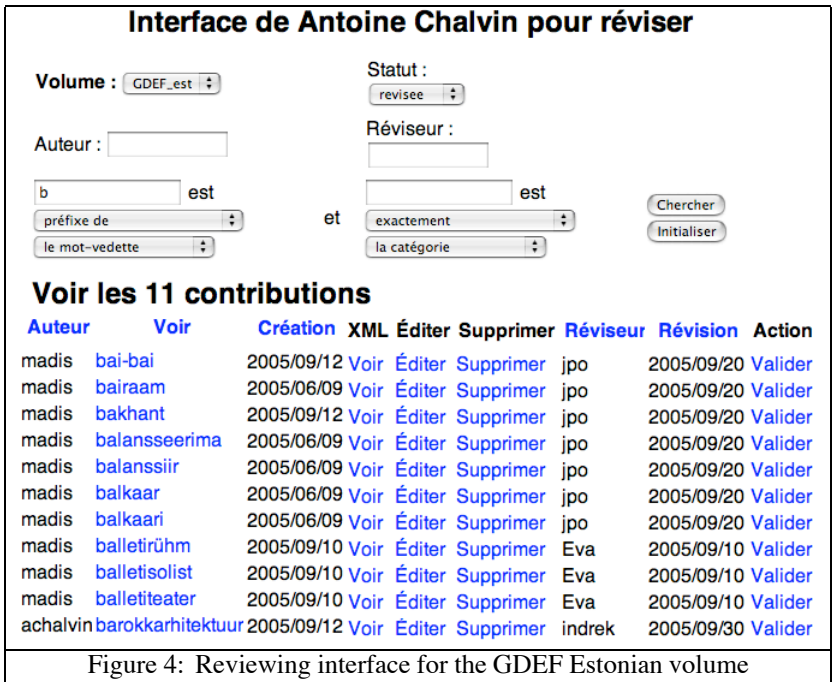


Figure 4: Reviewing interface for the GDEF Estonian volume

3. It is then validated by a validator;

When the entry is validated, it is integrated into the dictionary and all the users can search it. The reviewers and validators use the reviewing interface for searching entries and editing them (Figure 4).

6. Tasks Management

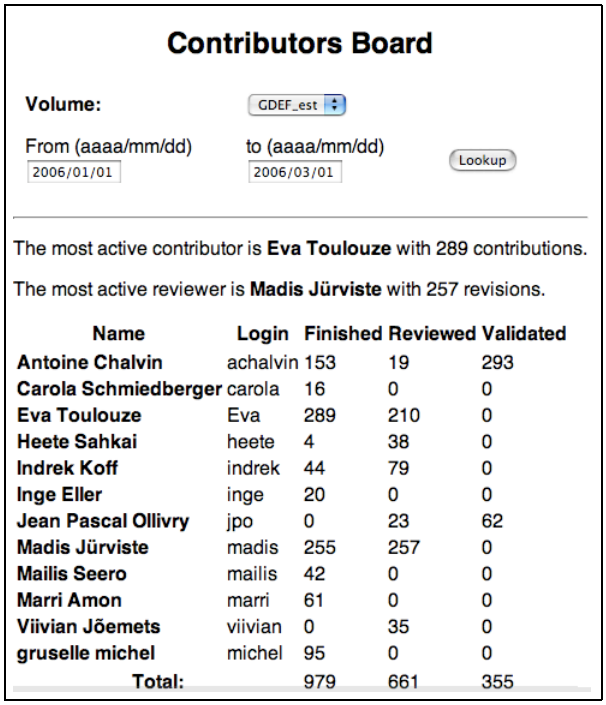
In order to manage the different tasks and roles, the platform gives the possibility to define groups and access rights. There are several groups with predefined rights:

- If the user is not logged, it can lookup the public re-

sources available on the platform.

- When the user is registered and logged, s/he is included de facto in the *contributors* group and can contribute through the entry writing interface.
- The users members of the *reviewers* group can revise the contributions written by users members of their working group.
- The users members of the *validators* group can validate the previously revised contributions.
- The last group is the one of the server *administrators*. They can manage users and their groups, add new resources on the platform, etc.

In the GDEF project, we added a supplementary constraint: the reviewers are divided into teams of two people: one with Estonian as a mother tongue and the other with French. Each one revise the contributions of the other one.
In order to facilitate the construction work of the dictionary and possibly the remuneration of the writers, it is possible to obtain a summary of all the contributions in a given period of time. Figure 5 shows the contributors board of the GDEF dictionary for January and February 2006.



Contributors Board

Volume: GDEF_est

From (aaaa/mm/dd) 2006/01/01 to (aaaa/mm/dd) 2006/03/01 Lookup

The most active contributor is **Eva Toulouze** with 289 contributions.

The most active reviewer is **Madis Jürviste** with 257 revisions.

Name	Login	Finished	Reviewed	Validated
Antoine Chalvin	achalvin	153	19	293
Carola Schmiedberger	carola	16	0	0
Eva Toulouze	Eva	289	210	0
Heete Sahkai	heete	4	38	0
Indrek Koff	indrek	44	79	0
Inge Eller	inge	20	0	0
Jean Pascal Ollivry	jpo	0	23	62
Madis Jürviste	madis	255	257	0
Mailis Seero	mailis	42	0	0
Marri Amon	marri	61	0	0
Viivian Jõemets	viivian	0	35	0
gruselle michel	michel	95	0	0
Total:		979	661	355

Figure 5: Contributors board for the GDEF dictionary

Then, in order to facilitate the revision of the entries, the dictionaries can be exported as a whole or by parts, in several formats (text, HTML, XML, PDF, etc.) and printed. After editing or reviewing offline the entries, the data can also be reimported into the database.

7. Conclusion

We presented here "Jibiki" platform, a generic dictionary online development platform that is used by several dictionary building projects such as the GDEF Estonian-French bilingual dictionary. This platform is based on research work in the domain. It is generic and can be used in (almost) any kind of dictionary development project from simple monolingual lexicons to complex multilingual pivot dictionaries as well as terminological resources.
The development of the platform is still an ongoing process. Once a stabilized version is obtained, it is planned to make it available for download freely as an open-source software. As of today, if you would like to use this platform for developing a dictionary free of rights, we invite you to contact us.

8. References

Antoine Chalvin and Mathieu Mangeot. 2006. Méthodes et outils pour la lexicographie bilingue en ligne : le cas du grand dictionnaire estonien-français. In *EURALEX 2006, à paraître*, Turin, Italie, 6-9 septembre.

Mathieu Mangeot and David Thevenin. 2004. Online generic editing of heterogeneous dictionary entries in papillon project. In *Proc. of the COLING 2004 conference*, volume 2, pages 1029--1035, Geneva, Switzerland, 26 August.

Mathieu Mangeot, Gilles Sérasset, and Mathieu Lafourcade. 2004. Construction collaborative d'une base lexicale multilingue. *Traitement Automatique des Langues*, 44(2):151--176, February.

Mathieu Mangeot. 2001. *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Septembre.

Gilles Sérasset. 1994. *Sublim : un système universel de bases lexicales multilingues et Nadia : sa spécialisation aux bases lexicales interlingues par acceptions*. Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1, Décembre.

Gilles Sérasset. 2004. A generic collaborative platform for multilingual lexical database development. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources Workshop*, pages 73--79, Geneva, Switzerland, 28 August.

Gilles Sérasset. 2005. Multilingual legal terminology on the jibiki platform: The lexalp project. In Mathieu Lafourcade, editor, *Proc. of Papillon 2005 Workshop*, pages 64--73, Chiang Rai, Thailand, 11-13 December.