

A standard based environment for the edition and distribution of multilingual terminological data: the DHYDRO platform.

Laurent Romary*, Jean-Luc Husson*, Marc Van Campenhoudt†,
Nadia Viscogliosi*

* LORIA (UMR 7503) Equipe Langue et Dialogue
BP239 - F-54506 Vandoeuvre-les-Nancy, France
{husson,romary,viscogli}@loria.fr

† Centre de recherche Termisti
Institut supérieur de traducteurs et interpretes
34, rue J. Hazard - B-1180 Brussels, Belgium
marc.van.campenhoudt@euronet.be

Abstract

The MLIS/Dhydro project aims to create an environment making it possible to manage and publish multilingual terminological data stemming from the computerization of the Hydrographic dictionary (published at present in three languages by the International Hydrographic Bureau - IHB).

We would like to set out the main features developed, and show the generic character of the choices made, namely through the use of a standard data format based on the Martif standard (ISO 12200) and expressed in XML (*eXtended Markup Language*). We will focus on the general editing process, which involves a large group of experts specialised in the field for each language used in the dictionary. They interact through a terminological forum (called « Log book ») and, thanks to a unified edition interface, can modify the terminological database diffused on the Internet.

We will present two complementary tools. Firstly, an interrogation interface designed to consult multilingual aligned texts – such as *The Law of the Sea* and its various official translations – that allows terminological occurrences in context to be observed and their equivalents to be identified. Secondly, a terminological database consultation tool that, thanks to XML and its associated stylesheet flexibility, offers various consultation modes, according to the user's needs (translation, teaching etc.).

We will conclude by presenting the results of a first evaluation carried out in collaboration with a group of users composed of people from the IHB and specialists in terminology who are not involved in the project.

1. Context and objectives

This paper is intended to provide a description of the functionalities of the Dhydro¹ platform for the editing and consultation of multilingual terminological databases. Initially, this platform was developed to allow the updating and computerized accessibility of the *International Hydrographic Dictionary* (IHD). Published by the International Hydrographic Bureau (IHB), this dictionary initially consisted of three monolingual volumes (English, French, and Spanish) containing hydrography-related terminology. These dictionaries were maintained by editors specialized in each of the languages. However, the geographical distance between them adversely affected the actual development of the dictionaries, whereas the printed format restricted the dissemination of terminology vital to the maritime community.

The Dhydro project has two major objectives:

- To provide the editors with the necessary communication tools to enable them effectively to interact with each other, to edit terminological data, to ensure quick access to the database and to publish it in various formats (bilingual or trilingual glossaries, monolingual dictionaries, etc.).
- To ensure the widest possible access to the terminological database, with the IHB having agreed to publicize the data resulting from the project².

While this clearly identified the thematic field, we also sought to ensure that the tools used would be independent of, on the one hand, any particular area of specialization, and, on the other, any specific IT platform. Intensive use was made of standards related to information technologies, data models, and encoding formats, whereas the aim was to create an editorial scenario that should be as coherent and as robust as possible. This has resulted in a highly flexible generic management platform which we have been able to test in close consultation with the end-users of the tool.

This paper does not aim to provide a full description of all the work that has been done by the participants in the course of the 18-month project since this is amply treated in another contribution to these Proceedings [1] which deals specifically with the conceptual model used and the retroconversion process of the three monolingual IHD volumes into a multilingual terminological database. The present paper will first focus on the editorial scenario

¹ DHYDRO (<http://www.loria.fr/projets/MLIS/DHYDRO/>) is a project within the European MLIS (Multilingual Information Society) programme. Its five partners are: the Bureau Hydrographique International (B.H.I., Monaco), the Laboratoire lorrain de recherche en informatique et ses applications (LORIA, Nancy), the Service hydrographique et océanographique de la marine (SHOM, Paris), the Centre de recherche TERMISTI (Institut supérieur de traducteurs et interprètes, Brussels) and the Institut für Deutsche Sprache (I.D.S., Mannheim).

² The aim is also to ensure a wide distribution of the program developed as a result of this project so that it may benefit other specialist communities.

that was used. Afterwards, we shall discuss the tools that were developed while closely examining the functionalities of the two essential components, i.e. the record editing tool, and the database consultation tool. The final part will be devoted to a brief presentation of the various technologies used.

2. The editorial scenario

The construction of an editorial scenario for the editing and consultation of the computerized terminology data involves defining a number of precise roles and responsibilities, and establishing rules for the management of multilingual records from their creation up until the validation. A high degree of rigor is indeed essential in a joint project involving the supply of distance-processed data.

2.1. The editorial board

The editorial board, which is responsible for the development of the terminological database, comprises the following three agents:

- **The president of the editorial board**

The president has a multiple role. First of all, he is responsible for managing all language-independent data, which are identified by the term «concept» within a given field of specialization, i.e. the area to which the concept belongs, the possible semantic links that connect the concept to other concepts in the database as well as the information illustrating the concept. The president is also responsible for channeling the work completed by the various editorial groups of each language and for the final validation of the contributions by the various editors. Finally, the president manages the database and decides on the addition of a new language and the appointment of new editors.

- **The editor**

The editor is responsible for a particular language, and is thus the only one who is authorized to carry out any changes in the conceptual records in this language. The limited number of editors responsible for a given language precludes problems related to editorial conflicts and competing access to data. The editor may alter semantic data (definition, the encyclopedic development, and any bibliographical references used) and lexical information (list of terms, links between the terms and attested usage contexts).

- **The accredited language advisor**

The editor responsible for a language may assemble a group of advisors to assist him/her in managing the language. In this capacity, they may use the same editing tool as the language editor, without however being entitled to export their records to the server. There are no limits as to the number of accredited advisors.

2.2. The databases

The system manages three databases:

- **A local database for each editor**

Each editor manages his/her own local database which contains all of the records on which he is working. As a result, the editing of the records may be done off-line.

- **A temporary database**

This database contains all of the conceptual records that are still in the processing stage, or that have not yet been validated by the president of the editorial board. In other words, this database provides an up-to-the-minute picture of the activities by the IHD editors.

- **A validated database**

This database only contains records that have been locally validated by each of the language editors, and by the president of the editorial board. This is the database from which derived documents (bilingual glossaries, monolingual dictionaries, etc.) will be drawn, at the request of the president of the editorial board. It is this database that internet users will be able to access through the consultation interface.

2.3. The life cycle of a conceptual record

It is possible to distinguish the following four stages in the life cycle of a conceptual record, as shown in figure 1.

2.3.1. Creating or importing conceptual records

Each editor is able to create a new record (phase 1bis). This is automatically identified as being unique, with the system allocating a default field (hydrography). Afterwards, only the president of the editorial board can modify this information, since this task involves work on the concept itself. In this phase, the new record is integrated into the language editor's local database of conceptual records. Another way of adding a conceptual record to this database is to import it (phase 1). The editing tool enables users to search the database on the server, and to select a set of records to be imported from among the replies. The search facility will be discussed below.

2.3.2. Off-line editing of the conceptual record

Once the record has been imported, the language editor may alter all lexical and semantic information for the language for which he is responsible (phase 2). The editing tool will also be described in detail below.

2.3.3. Exporting the conceptual record to the server

When the language editor is satisfied with the modifications and has validated them, the record can be exported to the server (phase 3). The record of the temporary database is updated on the server in order to

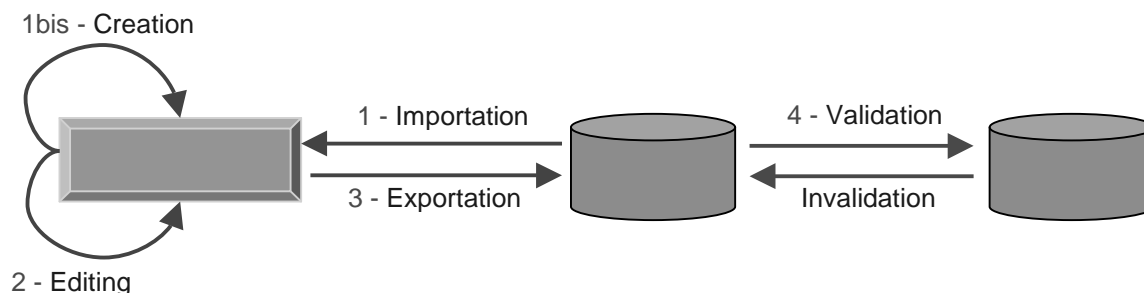


Fig 1: Life cycle of a conceptual record

integrate the modifications.

2.3.4. Validation of the record by the president of the editorial board

For the conceptual record to be included in the

from the Dhydro site. The main page lists all messages concerning all of the subjects for discussion introduced by the editors. These messages may be consulted according to three different sorting methods: in alphabetical order by author, then by date, date of submission, or topic. This

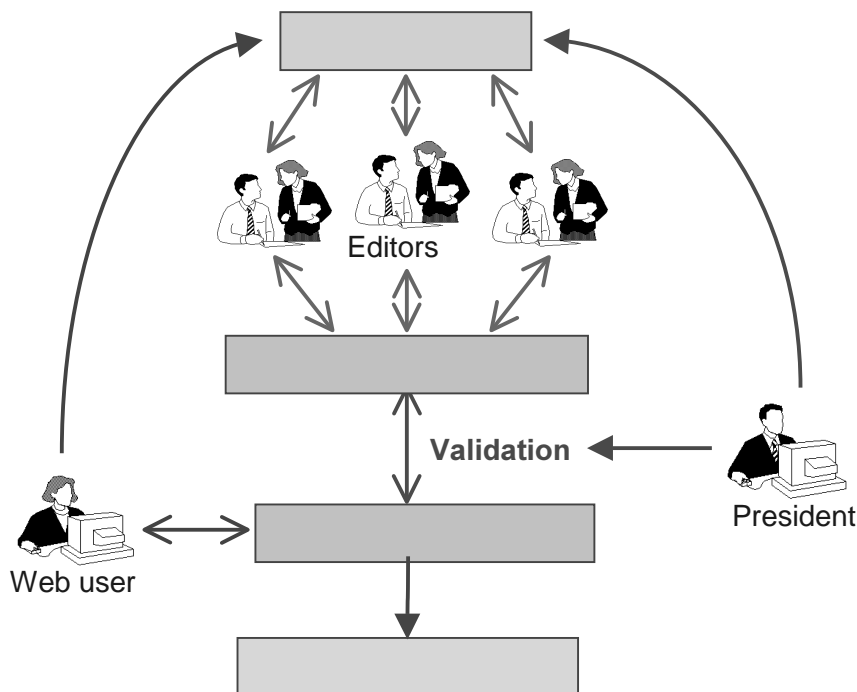


Fig. 2: Summary of interaction between actors

validated database on the server, it must first be validated by the president of the editorial board (phase 4), who may refuse the changes, or suggest to the other language editors to take them into consideration for their own language.

Figure 2 provides a summary of the actors and the interaction between them.

3. Tools

The platform contains four major tools: a forum-type joint communication tool (logbook), an editing tool for the editors and the president of the editorial board, a tool for the consultation of the database via the Internet, and a management and indexing tool for the database sited on the Dhydro server. All of these tools are accessed from one single site, called the editorial site, which is accessible only by IHD editors.

3.1. The logbook

The logbook constitutes an interactive communication space reserved for the editors of the IHD and the president of the editorial board. It consists of web pages accessible

page enables editors to introduce a subject for discussion, while messages can be accessed simply by selecting an existing message.

The page corresponding to the message provides the body of the text and enables the editor to reply to it in the logbook, or by e-mail to the sender.

3.2. The editing tool

This tool lies at the heart of the system since it enables access to the validated database, the importation of conceptual records from the Dhydro server, the off-line editing of local records, and the exportation of updated records. Although a description of the numerous functionalities of this tool falls well outside the scope of this paper, we should like to discuss some prominent aspects that have an immediate bearing on the conceptual model outlined in (Husson et al, 2000).

The main window of the editing tool (fig. 4) displays all the records included in the editor's local database. Each line corresponds to a conceptual record and provides the identification code, as well as the preferred term selected to represent the concept in each language.

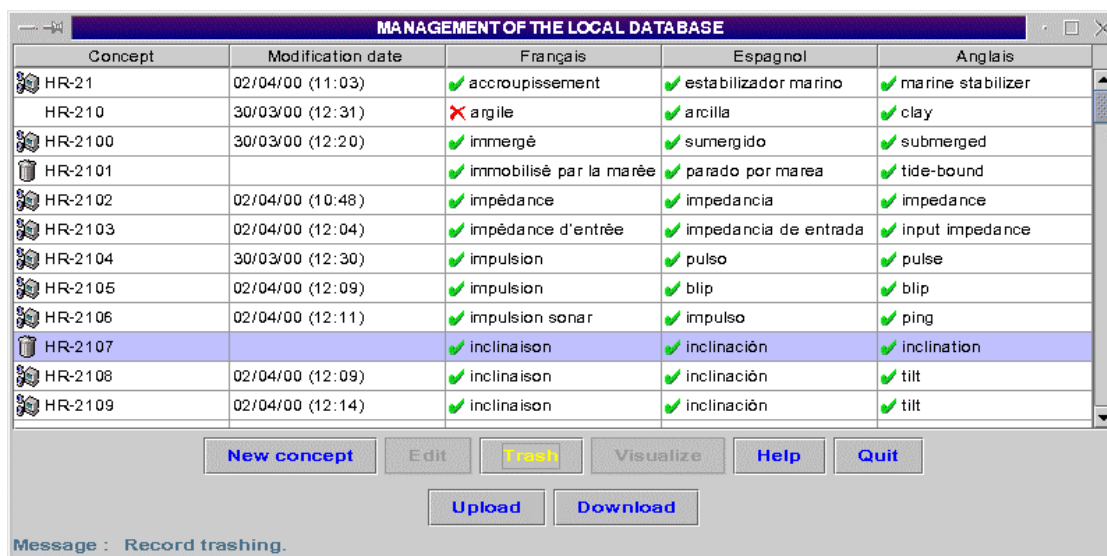


Fig. 3: Window for the management of a local database

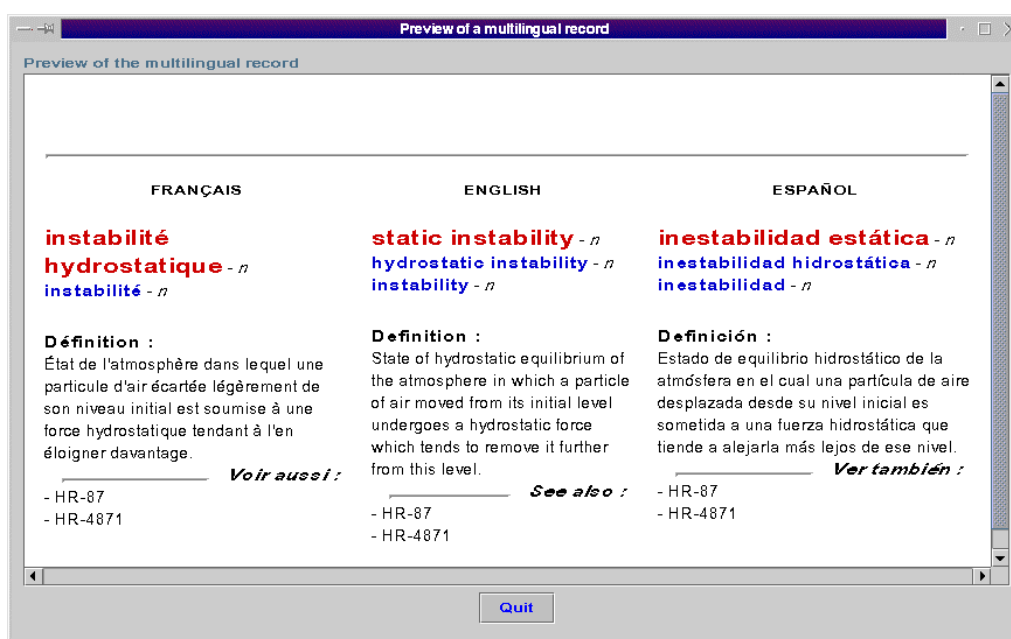


Fig. 4: Preview of a record

As a result of the split screen facility it is possible clearly to distinguish between the various types of information. The sample window below shows conceptual

data that can only be modified by the president of the editorial board (concept field, links between the concepts and the illustration).

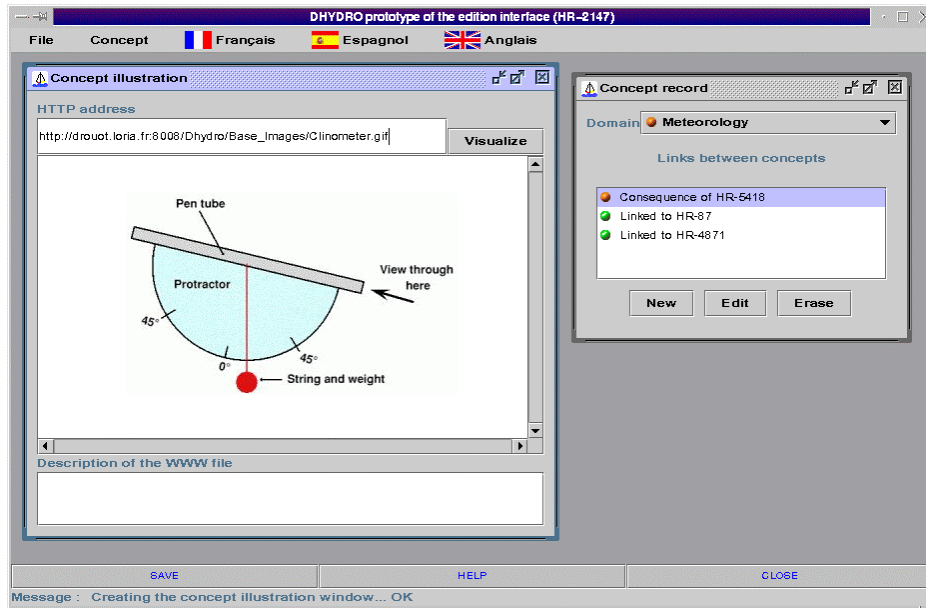


Fig. 5: Window for the management of conceptual data

When editing a record, the lexical and semantic information appears in separate windows. Figure 6 is an example of a record editing window and illustrates the presentation of lexical information in English (terms,

links between the terms and attested contexts of usage) and the list of descriptive fields for a given term.

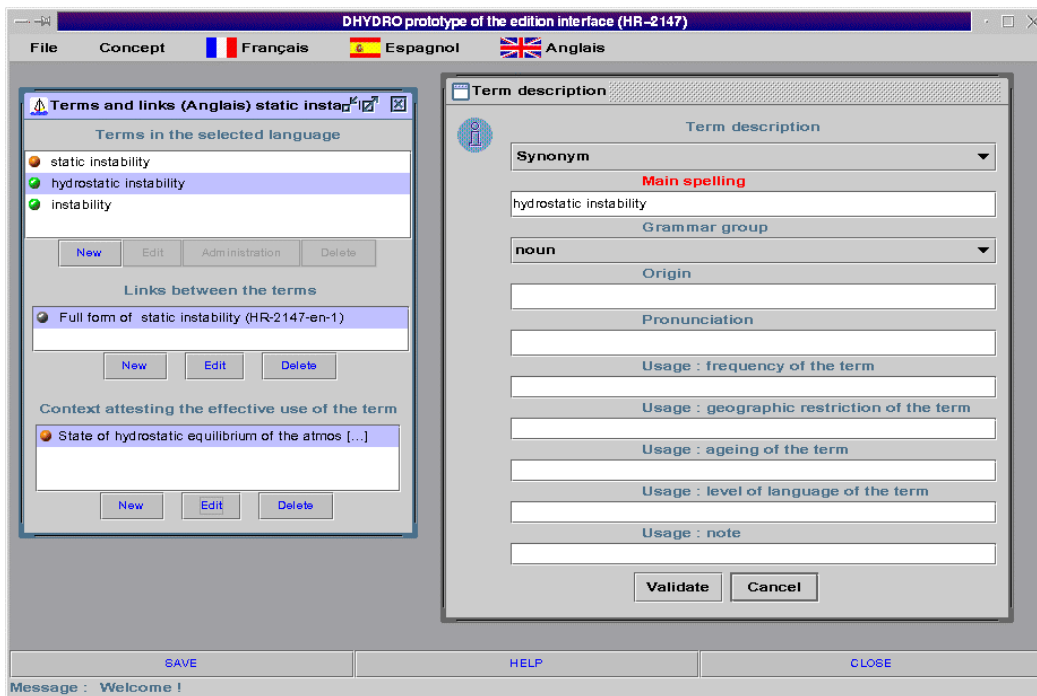


Fig. 6: Window for the management of lexical information

Figure 7 shows the windows for the modification of the English definition and the information related to the history of the changes made to this record.

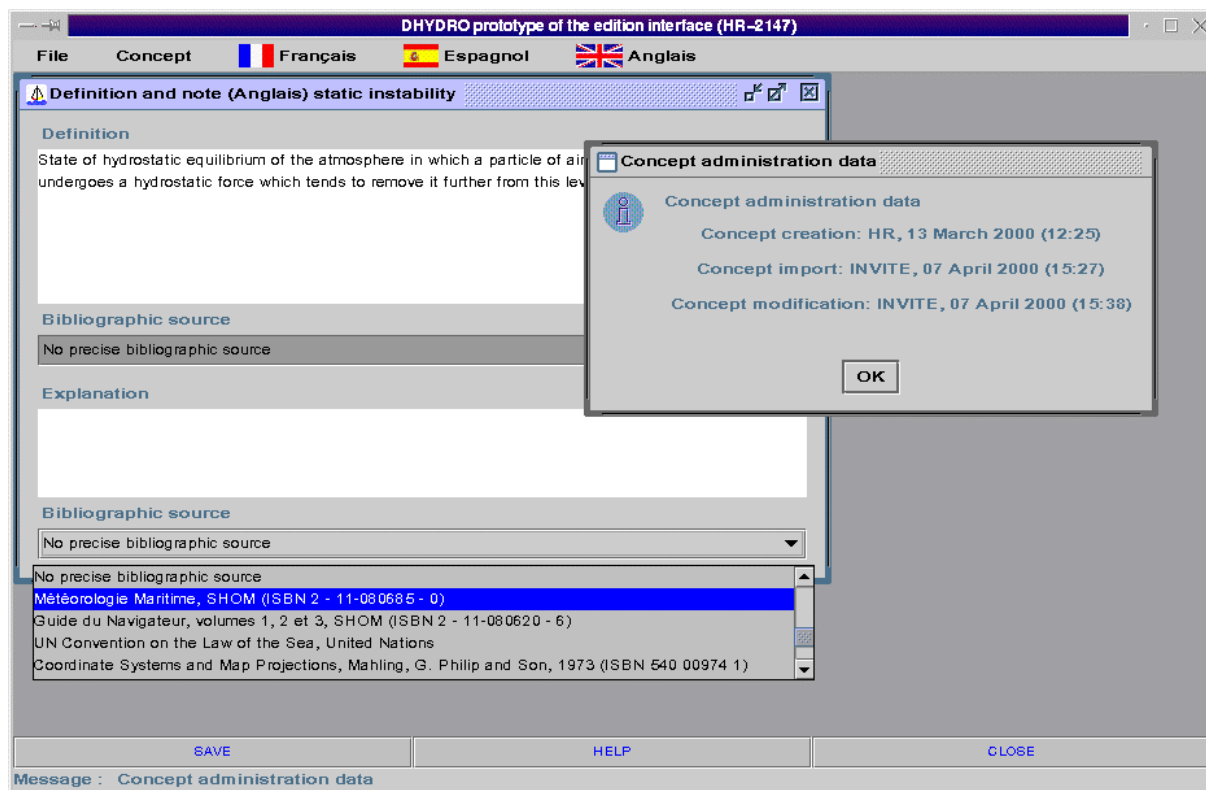


Fig. 7: Semantic information and administration history

3.3. The consultation tool

This tool is intended to allow each Internet user to access the IHD. At present, this tool is not yet available on the Internet, but the interface will be closely patterned on the search tool integrated into the editing tool. The ways of formulating requests will be identical, but in

accordance with IHB wishes not all users will be able to import or print out the records selected. The object is to offer only HTML access to the individual records. Figure 8 provides an example of a request to the IHD to search for a regular expression in the French terms. The search of the records that meet the request relies on an ad hoc indexing module implanted on the server.

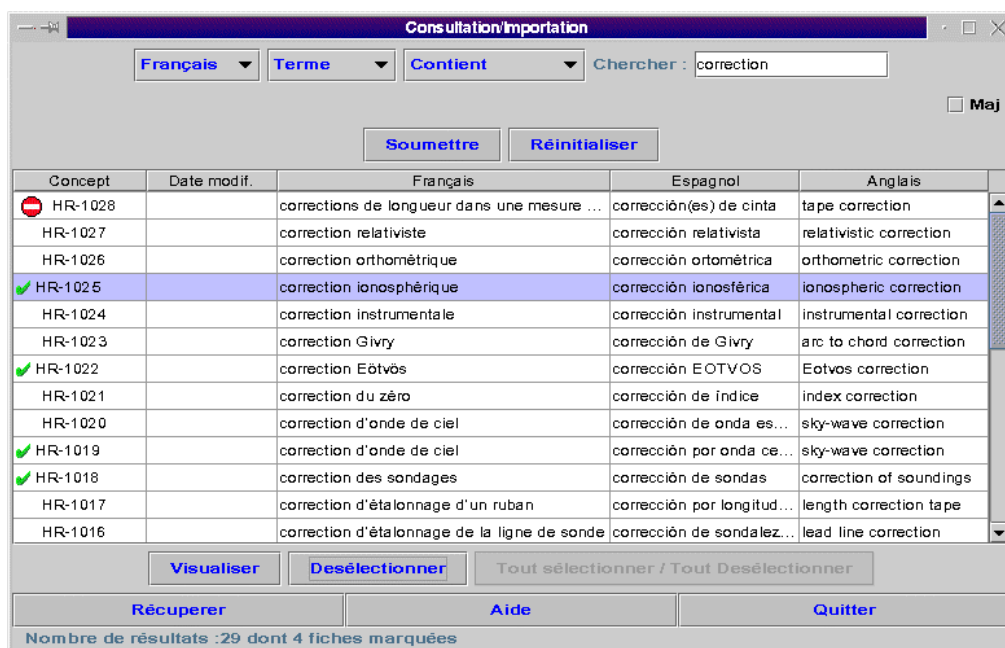


Fig. 8: Window for the consultation of the conceptual database

4. Overview of standardized technologies used

The use of the IT tools by LORIA, the French partner in charge of the software developments, is based on the choice of technologies that have been extensively tested in the course of other projects involving the processing of linguistic resources (Silfide and MLIS-Elan projects). These technologies may be divided into three categories, and are largely endorsed by European and/or international standardization bodies:

4.1. Structure and presentation of the documents

- The XML format (a simplified SGML, ISO 8879) for the encoding, structuring, and distribution of the electronic documents.
- The international ISO 12200 MARTIF standard. The MARTIF format is accompanied by a DTD (Document Type Definition, which expresses the abstract syntax of a document), specifying the XML coding of multilingual terminological databases with a view to enabling negotiated exchanges.
- The XSL proposition of W3C which is a language for the conversion of documents through a style-sheet facility – e.g. into HTML format destined for the consultation of the IHD on the Internet.
- James Clark's XT tool. This is a Java implementation of XSL enabling the application of style sheets to XML documents.

4.2. Management of network functionalities

- HTTP. HTTP was chosen because it offers a large number of high-end functionalities for the exchange of data within a client-server environment.
- The Nexus server. Designed in Java, this server operates on all material architectures and effectively implements the servlet technology that lies at the basis of our client-server architecture.

4.3. Development tools

- The programming language Java because of its portability which ensures the independence of applications vis-a-vis any material architecture.
- The SXP parser. This is a Java API that allows the processing of XML-encoded structured data.
- Hypermail . This is the tool on which the use of the Logbook depends. Hypermail enables the archiving of electronic messages and the HTML display of the archive, which can then be accessed on the Internet, in the form of a list of thematic subjects for discussion.

5. Conclusion and prospects

The use of communications networks for this customer-server architecture and specialized tools will considerably speed up the pace of development of the IHD. As a result, the periodic publication of the dictionary will make way for a continuous edition. The intensive use of both encoding and data processing norms and standards has enabled the creation of a generic and portable environment for the management of a multilingual terminological database.

There are three phases in the development of the tasks to be completed.

- First of all, it is important to assess both the proposed editorial scenario, and the robustness of the tools developed within the specific framework of the Dhydro project. Our partners in the IHB will have a crucial role to play in this respect, and the success of this experimental stage will undoubtedly depend on the further development of this project.
- If the outcome is deemed satisfactory, the second phase will involve the addition to the current platform of a user-friendly server management tool so that these tools can be made easily available to the wider community of terminologists, interpreters, and linguists with a view to enabling them independently to construct multilingual terminological databases devoted to various specialized fields.
- Finally, it is hoped that the development of these individual terminological databases will be conducive to the construction of a unique network in which all these data would be structured. There are numerous applications in such areas like knowledge management or automatic language processing that make use of this linguistic resource tool.

6. References

- International Hydrographic Organization (1994): Hydrographic Dictionary, 5th ed., Monaco, Bureau hydrographique international.
- ISO 8879 (1986): Information processing - Text and office systems - Standard Generalized Markup Language (SGML) , Geneva, International Organisation for Standardisation.
- ISO 12200 (1999): Computer applications in terminology -- Machine-readable terminology interchange format (MARTIF) - Negotiated interchange, Geneva, International Organisation for Standardisation.
- LORIA & centre de recherche TERMISTI, 1999: Le modele éditorial du forum Dhydro, deliverable D1.1 of the MLIS-2009 Dhydro project, March 1999.
- LORIA & centre de recherche TERMISTI, 1999: Choix techniques d'accès aux données, échanges et outils, deliverable D2.1 of the MLIS-2009 Dhydro project, March 1999.
- Organisation Hydrographique Internationale (1998): Dictionnaire hydrographique, 5th ed., Monaco, Bureau hydrographique international.
- Organización Hidrográfica Internacional (1996): Diccionario Hidrográfico, 5th ed., Monaco, Bureau hydrographique international.
- Husson Jean-Luc, Nadia Viscogliosi, Laurent Romary, Sylviane Descotte and Marc Van Campenhoudt 2000. From specialized lexicography to conceptual databases: which format for a multilingual maritime dictionary? Second Conference On Maritime Terminology, Turku, Finland, 11-12 May 2000.