XTERM: A FLEXIBLE STANDARD-COMPLIANT XML-BASED TERMBASE MANAGEMENT SYSTEM

Lorenzo Piccioni, Eros Zanchetta

SSLMIT – University of Bologna
Corso della Repubblica, 136
47100 – Forlì (FC)
Italy
lpiccio@sslmit.unibo.it
eros@sslmit.unibo.it

Abstract

This paper introduces XTerm, a Termbase management system (TBMS) currently under development at the Terminology Center of the School for Interpreters and Translators of the University of Bologna. The system is designed to be ISO and XML compliant and to provide a friendly environment for the insertion and visualization of terminological data. It is also open to the future evolution of international standards since it does not rely on a closed set of hard-coded data representation models.
In this paper we will first introduce the project "Languages and Productive Activities", then we will outline the main features of the XTerm TBMS: XTerm.NET, the graphical user interface (the main tool of the terminographer), XTerm.portal, the web application that provides online access to the termbase and two tools that provide innovative functionalities to the whole system: CARMA and COSY Generator.

1. The project "Languages and Productive Activities"

The project "Languages and Productive Activities" has been started in november 1996 with the goal of promoting collaboration between the University of Bologna and companies located in Emilia Romagna and surrounding regions through the creation of a termbase accessible to students, interpreters, translators and terminologists.
The XTerm Termbase Management System has been designed to reorganise terminological data resulting from more than 7 years of work carried out by students, researchers and staff of the School for Interpreters and Translators within the framework of the project.
Almost every entry in the current termbase (http://www.terminologia.it) has been revised by in-house experts working in one of the many companies that participate in the project (e.g. Ferrari, Aprilia and Ferragamo).

1.1 – The current termbase

All of the current data has been either produced or processed using SSLiMIT Trad, a proprietary application developed in 2000 targeted at the needs of the project.
When the work on the creation of the unified termbase started we were aware that the terminological data we were processing had a major shortcoming – terminological data in the termbase do not comply to the specifications of ISO 12200 (Computer applications in terminology -- Machine-readable terminology interchange format (MARTIF) -- Negotiated interchange) and ISO 12620 (Computer applications in terminology, Data categories) – but as the work progressed and the number of entries grew it became evident that the system currently being used also had other flaws:

1. the work of every terminographer is treated as an independent collection of data, unrelated to any other entry in the termbase;
2. as a consequence of this, terminological data belonging to the same domain may be unrelated if the work has been carried out by more than one terminographer;
3. different terminographers used slightly different metalanguages in their work, leading to inconsistencies that had to be eliminated with a time-consuming revision;

These flaws have two negative consequences: data inconsistency and waste of resources.

1.1.2 – Data inconsistency

When we started putting together the terminological data collected through the years we expected to find quite a few duplicated terms. As the work progressed, however, it became evident that, on a few occasions, duplicated terms belonging to the very same domain had a different definition (Zanchetta, 2003).
We soon realised that given the sparse nature of the termbase, it would be extremely difficult and time-consuming to keep track of duplicated data and prevent inconsistencies such as these.

1.1.3 – Waste of resources

Apart from requiring a considerable revision effort, the current system relies heavily upon technical support and manual editing of the databases.
Moreover, in order to be made available online, data currently has to be entirely reprocessed and converted to a different database format (i.e. from Microsoft Access to mySQL).

Needless to say, this process tends to be long and cumbersome.

1.2 – Desiderata for the new termbase

XTerm is an attempt to address all these problems and provide a solution to achieve the two goals of the project: that of creating a scientifically grounded termbase and that of producing linguistic data usable by our partner companies.

1.2.2 – ISO compliance

Once the data in our termbase has been revised, converting it to an ISO compliant XML format is a quite straightforward process. The challenge is another: ISO 12200 and ISO 12620 standards are undergoing revision, but even if they were not, we still would not want to create a closed system: the new system has to be flexible enough to be able to take into account every possible change to the standards and even the creation of new ones.
This has been achieved by adopting a mechanism that we named "Metamorphic terminological data definition" (see below).

1.2.3 – Productivity and flexibility

XTerm has been designed with productivity in mind: terminographers need a tool that simplifies their work, something they can learn to use in just a few hours and that allows them to quickly compile a great quantity of terminological data.
The web interface also reflects this philosophy: users can choose the amount of information to visualise, those interested in the knowledge base will find all the details they need, others may choose to view just a bare list of terms with their equivalents in various languages.

1.3 – Normalization and migration

At the time of writing (late February 2004), normalization of existent terminological data is still underway, all terminological records will be converted to the most recent SSLiMIT Trad format.
Once the new system is complete, the data will be migrated to the new database format.
The project has so far produced more than 70.000 terms and the number is bound to increase as new students graduate in terminology.
The final result of the upgrade to XTerm will be a constantly expanding termbase that needs virtually no technical support and whose data can be seamlessly exported to and imported from any XML terminological interchange format.

2. XTerm

XTerm has been created within the framework of the "Project Languages and Productive Activities" but it not limited to it. In fact it can easily be adapted to different terminological projects since it is not merely a termbase, it is a TermBase Management System (TBMS).

The whole system consists of 4 main components:

- the database engine (mySQL, Oracle, Access) which takes care of the low level handling of the raw data;
- Xterm.NET, a graphic environment for data insertion, termbase management, querying and visualisation;
- one or more XML configuration files defining the data structure of the termbase (a virtually unlimited number of differently structured termbases can be hosted on the same machine);
- XTerm.portal, a web application that provides general access to the termbase(s) through a comprehensive querying engine;

2.1 – XML Terminology for Networks

XTerm.NET is a terminology management solution that allows users to create, manage and view multilingual terminological databases.
Since it is Unicode compliant, XTerm.NET is capable of managing anything from a small monolingual project to a great number of large projects containing millions of terms in all ISO 639 defined languages (depending on the capabilities of the underlying database engine[1]).
The system combines the data-consistency of traditional relational databases with the flexibility of the XML Schema definition. It becomes thus not only possible but also extremely easy to customise and replicate the terminological database structure.
The application consists of a small core of base projects and data-handling functionalities. Such a structure is then expanded via a number of dedicated plug-ins (i.e. small programs that perform very specific tasks) resulting in a highly modular and open system.
Plug-ins developed so far include:

1. database-related plug-ins (used to connect the application to various database engines);
2. visualisation-related plug-ins (useful to customise the rendering of visual information);
3. import and export plug-ins (that allow terminological data exchange and conversions between XTerm and almost every other terminological format);
4. CARMA (still under developement, helps the terminographer in keeping track of relations among terms in the termbase)

The XTerm.NET interface is meant to be extremely user-friendly. It is graphically integrated with intuitive icons and toolbars to ensure smooth navigation as one work with the application.
The uniform look-and-feel is designed to reduce learning curve, through consistent use of the latest Microsoft Windows system standardised features.
A highly customizable interface allows users to choose their own personal settings making the working environment more "comfortable"
The application is also to include an easy-to-follow, indexed, online help with a built-in search function that

1 At the time of writing our main database engine Unicode support is only available in the 4.1 alpha release of our main database engine (mySQL), we expect a final version 4.1 of the DBMS within the next few months.

will assist the user in finding information on any XTerm.NET feature or function.

2.2 – Metamorphic terminological data definition

The increasing need to experiment with different terminological data memorisation and representation schemes is dealt with in XTerm by adopting a standard data definition language: XML.
Starting from data definitions stated in the flexible but well-defined XML format, XTerm.NET encapsulates the hierarchical structure defined by an XML configuration file in a relational data definition representation that it is able to handle no matter what the data definition itself contains or consists of; the system will accept just about *any* data structure that conforms to a minimum set of limitations imposed on the XML file by an XML Schema, making the whole system truly dynamic and easily adaptable to any update to the fast-changing terminological interchange formats.

2.3 – Scaling and portability

The XTerm.NET terminology system can be scaled up to adapt to the needs of large companies/institutions that want to utilise a central database server (currently, we support mySQL and soon we plan to add support for Oracle Sybase) in order to allow a large number of users to work on a single terminology system, and at the same time can be scaled down to adapt to the needs of freelance translators using a desktop database (such as Microsoft Access).
XTerm.NET's ability to communicate with different database engines is currently achieved by a set of dedicated plug-ins.
We are planning to develop a web service component that will make XTerm totally independent from the underlying database engine and the graphical client when using a remote connection to a central database server.

2.4 – ISO 12620 compliance

XTerm is being developed with XML and ISO in mind, to ensure maximum compatibility with other termbases.
The termbase currently under development has been defined as a superset of ISO 12620 (i.e. it contains a higher degree of specificity, especially as relations are concerned).
The reason why we choose a superset of ISO 12620 is twofold: firstly the need to experiment and therefore ensure that researchers have all the instruments they need. Secondly the fact that companies may require that additional, non-standard compliant information be included in the database.
By using a superset of ISO, the system is capable of integrating all these additional features while maintaining downward compatibility with international standards.
Thanks to XML support, terminological data can then be exported in standard formats such as TBX and MARTIF. Moreover, the metamorphic data structure ensures compatibility with new terminology interchange formats and with future updates of the current standards.

2.5 – XHTML 1.1 compliant web interface

Web-based access to the termbase is achieved through XTerm.portal, a web application that interacts directly with the termbase.
In the ideal implementation of the system, new terms will be online in real time, that is as soon as they are inserted in the termbase.
Terminographers will work directly on the termbase and the result of their work will be immediately available on the Internet without mediation.
The portal provides a wide array of search and visualisation options to suit the needs and tastes of every user.
"One-button search" provides an extremely simple search mechanism while "Expert search" allows users to fine-tune the search options to retrieve accurate results.
Terminological record display can be customised to show the type and quantity of information the user desires: from simple bilingual glossaries to full-fledged terminological entries containing linguistic, semantic and ontological information (such as grammatical notes, contexts, definitions and relations).
XHTML 1.1 compliance ensures cross-platform compatibility and accessibility as well as portability to a broad range of client devices.[2]

2.6 – CARMA (Computer Aided Relation MAnager)

Computer Aided Relation MAnager is a plug-in that helps the terminographer in establishing relations among entries in the termbase and in then keeping track of them.
Ideally every entry in the termbase has to be related to other entries, CARMA helps in keeping relations consistent throughout the database and avoiding broken links in the conceptual systems by suggesting possible relations among terms.
This is done by analysing existent relations and proposing new ones (e.g., if A is related to B, and B is related to C then it is likely that A is also related to C).
Inferability rules are specified in the XML configuration file that defines the structure of the termbase and are therefore completely customisable.

2.7 – COSY Generator (COnceptual SYstem Generator)

This plug-in generates graphical representations of conceptual systems automatically or semiautomatically.
The relations expressed among terms in the termbase are used to generate visual representations of the conceptual systems, which terminographers can then edit to suit their own tastes.
COSY represents a major productivity improvement since the terminographer who wants to provide graphical representations of the conceptual system is no longer forced to manually draw them, possibly wasting even more time in learning to use a drawing tool.

2 At the time of writing XHTML 1.1's features are not yet fully supported by mainstream web browsers.

3 – Conclusions

In this paper we introduced the project "Languages and Productive Activities", we discussed its goals and features and described the work that has been carried out by student, staff and companies that contibuted to it.
We then outlined the limitations that became evident during the creation of a termbase collecting all the terminological data produced within the project.
Finally we proposed the creation of XTerm as a possible solution to the current limitations.
At the time of writing XTerm is a prototype in its early development stage. The system is not yet fully functional and migration of existent data has not begun yet.
As of the time of writing there are as yet no specific plans for its release policy. However, we do not rule out the possibility to make it available for research purposes to other universities and institutions in forms yet to be defined.

References

Zanchetta, E. (2003). Presentazione di un prototipo di Termbase Management System flessibile e standardizzato basato su XML. Dissertation: SSLMIT, University of Bologna.