

MLIS-5016 TransAccount WP1.1

Knowledge Base Design

Eurosources SA
27 March 2000

This document describes the architecture of the knowledge base. It gives an overview of data storage formats, access mechanisms, basic relationships between the different sections of the knowledge base, search, browsing and editing mechanisms. This document will be complemented with separate detailed design documents for each section of the knowledge base.

1 Overview

The Knowledge Base includes 4 sections (see Figure 1)

1. the *Term Database* including the appropriate English and French financial terminology,
2. the *Ontology* describing accounting concepts and their relationships, and
3. the *Translation Memory* including a set of aligned translated texts from annual accounts,
4. the *Reference Database* including a set of reference documents (legal, practice) on annual accounts.

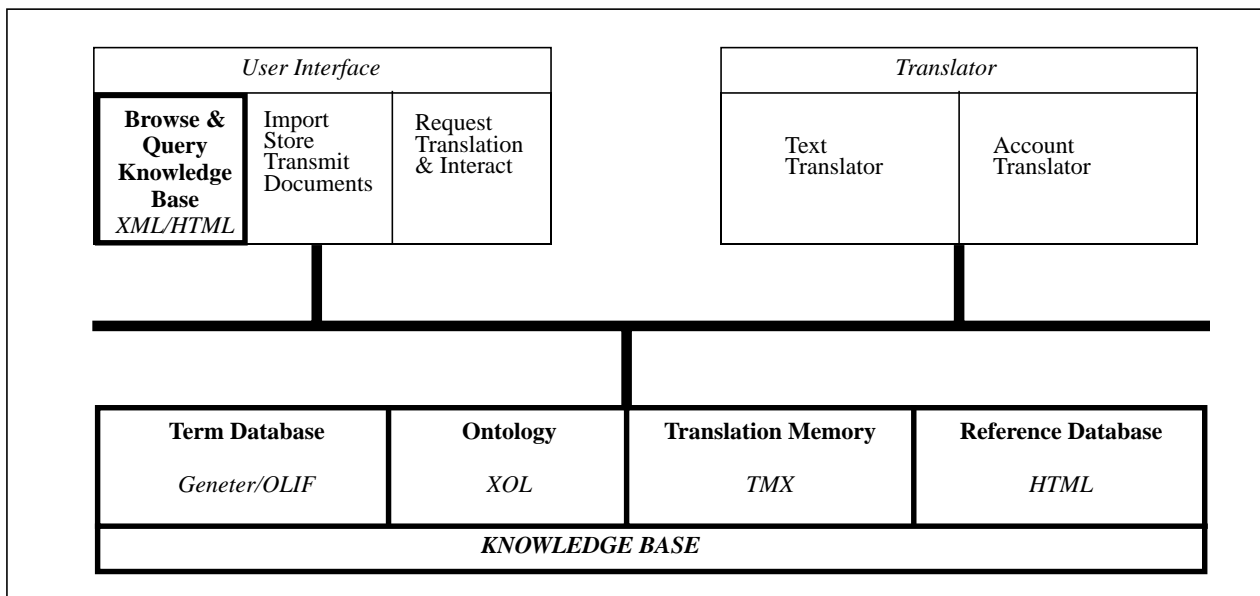


Figure 1: TransAccount System.

Each section of the knowledge base will be represented as a set of XML documents and will use existing XML DTDs:

- The Term Database will use Geneter extended using Olif (Olif is a non-XML standard develop within the Otelo project (<http://www.linglink.lu/hlt/projects/otelo/>) for representing machine translation dictionaries and term banks.
- The Ontology will use XOL (Ontology Exchange Language: <http://www.oasis-open.org/cover/xol.html>).
- The Translation Memory will use TMX (<http://www.lisa.org/tmx/index.html>)
- The Reference Database will use XHTML (XML-conformant HTML).

These choices determine most of the architecture and implementation of the Knowledge Base. In order to focus on developing the content of the various section and to minimize the time spent on software development, we will use a number of simple XML toolsets and adopt simple storing mechanisms:

- All XML files will be stored as files. Since the data format is already defined, this leaves only the directory structure to specify and much of this directory structure will depend on the XML toolset supporting the implementation of the Knowledge Base.
- XML files will be index using XML indexing and search toolsets (several are freely available) to support full-text and structured queries.
- Browsing will be supported either by available XML browsers and/or by HTML browsers with XML-to-HTML conversion. Navigation between sections of the Knowledge Base will be supported by the XML Xlink mechanism.
- Editing will be supported by either generic XML editors or by specific form-based editors.

Additionally, structured indexes will be provided for some of the sections of the Knowledge Base. Navigation and cross-references will be supported by XML linking mechanisms. However, there are currently few implementations supporting XML links and we might have to develop a specific linking mechanism for cross-referencing elements in different sections of the Knowledge Base.

Therefore, we envisage a classical 3-tiered model, depicted in Figure 2, where each section of the Knowledge Base is stored as XML documents and accessed through a specialized server. The Knowledge Base server provides an integrated view of the different sections of the Knowledge Base. It analyzes a user request and dispatch the sub-requests

to each specialized server, and recombines the results into a single XML document transmitted to the user. The Knowledge Base server also implements a cross-referencing mechanism supported by external XML link database.

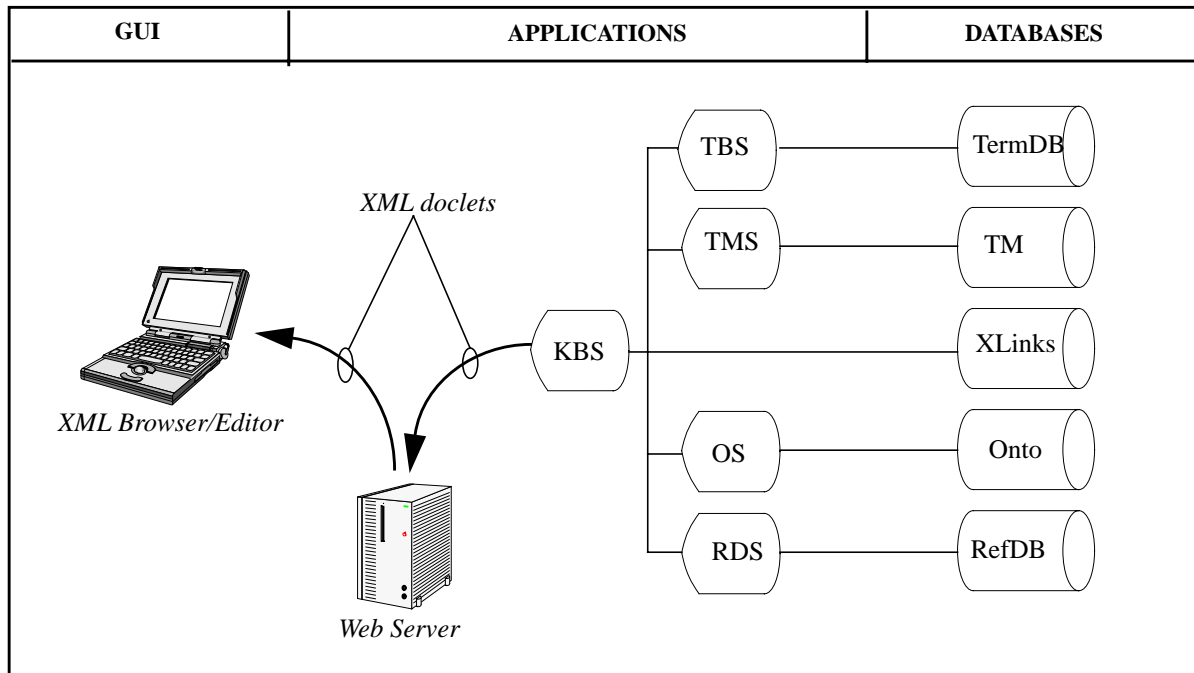


Figure 2: Knowledge Base Architecture.

2 Knowledge Base Servers

The various sections of the knowledge base will be encoded in XML using the appropriate DTD (XSL is not yet supported by existing XML tools)). This choice enables the use of existing XML standards for the representation of various sections, and the use of publicly available XML tools (parsers, browsers, editors), and will facilitate the exchange of sections of the knowledge base between partners. Until the new XML-compatible version of HTML (XHTML) is available, the Reference Database will be stored as a set of (linked) HTML documents.

For exchange purposes, the various sections of the Knowledge Base will be provided as simple XML files. The following section describes the architecture of the Knowledge Base in the context of browsing and editing.

2.1 Indexing and Search

For the purpose of browsing and editing, each section of the knowledge base will be associated to one or more indexes allowing for retrieval of content elements. The XML databases will be indexed using a specialized XML indexing tool to enable full text search as well as structured queries.

Access to a content element is provided through an index and a search procedure (using prefix or exact search) returns a list of matching keys and for each key a list of strings containing XML text. The list of keys can be used to build a clickable list of retrieved elements as in standard HTML search engines. Each XML element ('doclet') can be parsed using a standard XML parser, and further manipulated and processed using the XML object returned by the parser. Each section of the Knowledge Base will have a primary index used for basic navigation. Some sections, such as the Terminological Database may have more indexes, to enable search on several elements (conjunction or disjunction) of a term entry for example.

2.2 Linking and Navigation

The elements within a Knowledge Base section will be linked to provide clickable navigation links. Links will be formalized as XML links and translated as HTML links for the purpose of HTML browsing (by the XML-HTML translators) if there is no conveniently available XML browser.

Some links will be specified as part of some XML documents (in-line links). For example, translation links will relate English and French lexicons; a list of synonyms in a term entry will contain links to the appropriate synonym entries. In-line links will also define cross-references between different sections of the Knowledge Base. For example, a term entry which semantic zone is described by a concept name will be linked to the Ontology through that concept. The user clicking on the concept name in a term entry will for example bring up a new browser window displaying the concept of the Ontology, its definitions, and its links to other concepts, as well as links back to all the terms associated to that concept in both languages.

Some other links will be extracted automatically by cross-referencing elements in different sections. For example, in the reference database, occurrences of terms defined in the Term Database can be automatically linked to the relevant Term Base entry. These links will be stored externally as out-line links to avoid to modify the Knowledge Base.

Instead of direct links, some elements may generate a search query in another section of the Knowledge Base. For example, the headword of a term entry may be used to search the Translation Memory or the Reference Database for occurrences on this term.

Finally, sections of the database, such as the Reference Database and the Ontology, will also be browsed through hierarchical indexes or Tables of Content (TOC). These documents will be either prepared manually (TOC) or automatically (indexes) and linked to the relevant documents.

2.3 Access, Display and Editing

Depending on the maturity and availability of XML technology, browsing will be implemented using:

- XML browsers and XML editors
- XML-to-HTML transducers for display using HTML browsers, or using DSSSL (or XSS) for display in XML browsers. The current choice is to use an XML-to-HTML transduction scheme.
- Specialized editors implemented as HTML forms or applets.

For on-site development, a simple editing function can be provided using available XML or HTML editors, or a combination of text editors and XML parsers for validation. Depending on the maturity of XML technology, more advanced editing functions will either be implemented in an ad-hoc fashion or using generic XML editors.

3 Content of the Knowledge Base

3.1 Term Database

The Term Database format will be based on the Geneter XML format modified to cover the Olif specification. The DTD may be completed to support additional elements not covered in the standard. These additional elements may be mapped to standard elements for exchange purposes. The exact list of features and values for the Term Database will be specified as part of the DTD extension.

A term entry will be linked to the Ontology through the semantic zone describing the conceptual structure of the term; the main concept will be used to generate a direct link to the description of this concept in the Ontology.

A term will be linked to the Translation Memory through its primary key (the headword or citation form), and possibly through elements of its definition: selecting the key for search in the Translation Memory will call a Translation Memory browser window with the headword as the search term.

A term entry will also be linked to the Reference Database either through a definition link or through a search query in the same way as for the Translation Memory.

3.2 Translation Memory

The Translation Memory will be encoded using the TMX XML standard. The basic search mechanism of the Translation Memory is a KWIC search performed on the source or the target side of the Translation Memory. The query term is a string. More advance search facilities involve structured XML queries.

3.3 Ontology

The Ontology will be encoded using the XOL standard. The Ontology will be indexed on concepts and relations for search on concept names and on relation names. Relations will enable linking between concepts. Additionally, each concept will be linked (through concept indexes) to related language specific terms. XOL will be extended where necessary to support arithmetic constraints for modelling account elements.

3.4 Reference Database

The Reference Database will be encoded using XHTML and accessible through a full text index for full text search, a Table of Content, and an automatically generated clickable index. Links will be provided from the Table Of Content and the clickable index and the Terminological Database.

4 References

- Olif: <http://www.linglink.lu/hlt/projects/otelo/>
- TMX: <http://www.lisa.org/tmx/>
- XML links: <http://www.w3.org/TR/xlink/>
- XOL: <http://www.oasis-open.org/cover/xol.html>
- XQL: <http://metalab.unc.edu/xql/>