

Sharing Knowledge Across Language Barriers: A Universal Approach for Online Books.

Prof. Magdy Nagi
Bibliotheca Alexandrina, P.O. Box 138,
21526, El Shatby, Alexandria, Egypt.
magdy.nagi@bibalex.org

Prof. Tarcisio Della Senta
UNDL Foundation, 48 Route de Chancy
CH-1213 Petit-Lancy, Geneva,
Switzerland
tdellasenta@undlfdoundation.org

Abstract

Our focus is on a real case of a massive collection of online books (e-books), the case of EOLSS (Encyclopedia of Life Support Systems). It is a massive collection of documentation, under constant change, aiming at different categories of readers coming from multiple linguistic and cultural backgrounds. The current paper aims at presenting a system to navigate bibliographical information of books as well as their contents in any natural language. The system has three modules. The first module deals with the semantic representation of the content of the book using UNL technology. The second module deals with representation of cataloging information. The third module deals with the interface through which the user will read both of the bibliographical information and the content of a given book in his/her native language.

ACM Categories & Subject Descriptors:

J.7 Computer Applications, COMPUTERS IN OTHER SYSTEMS, Publishing

General Terms: Languages

Keywords: Encyclopedia of Life Support Systems, Library information systems, Machine translation, Universal networking language, eBooks

1. The UNL Technology: empowering multilingual interactive communication

EOLSS is an Encyclopedia made of a collection of 20 encyclopedias, online, and soon in the form of e-books [1]. EOLSS is unique in:

- size: an unprecedented global effort over the last ten years, with contributions from more than six thousands scholars from over 100 countries, and edited by nearly 300 subject experts. The result is
- “a virtual library equivalent to 200 volumes, or about 123,000 printed pages ... that is continuously augmented and updated.”
- its goal is to provide a firm knowledge base for future activities to prolong the lifetime of the human race in a hospitable environment.

The concern for UNESCO and for the EOLSS authors is to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BooksOnline'08, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558-249-8/08/10...\$5.00.

make EOLSS available to as many nations as possible, while expanding and updating its English content continuously. Translating it in every possible language is a daunting task that requires years of work and large amount of human and financial resources, if done in the conventional ways of translation.

Our proposal is to use the UNL System for representing both of cataloging data and the content of books in terms of language independent semantic graphs. Once this has been achieved, the language independent graphs can be decoded into any natural language, which in turn will give a semantic translation of EOLSS documents into multiple languages. Work has actually started with the six official languages of UNESCO. With the UNL System, this can be achieved in a relative shorter period of time, and at lower costs in comparison to costs of traditional translation.

2. The UNL Technology: empowering online Knowledge access and sharing

2.1 The UNL system

One way of introducing the UNL is to present it as the “language of computers”, which is different in nature from the concept of “computer language” such as JAVA, BASIC, C++ etc. Here, “UNL Language” is taken in the same meaning as when we refer to the “human languages”. In fact, UNL has lexical, syntactical and semantic components as does any natural language. It can therefore, represent all information, data and knowledge that humans produce in their own natural languages. The basic difference between UNL and human languages is that UNL is written in machine readable format. It represents information and data in a “digital alphabet” that machines can “understand” and process.

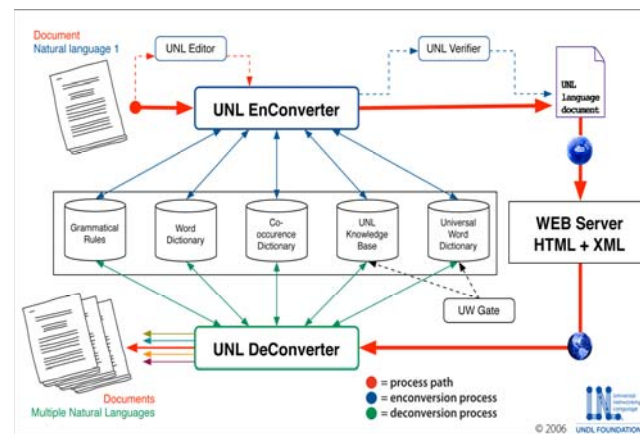


Figure 1: The core architecture of the UNL system

The UNL is the core of the UNL System, which comprises three sets of components [2]:

1. Linguistic components: dictionaries that include Universal Words (UWs) and their equivalents in natural languages, grammatical rules responsible for producing a well formed sentence in the target natural language and knowledge base responsible for representing a universal hierarchy of concepts in natural languages;
2. Software components: two software programs for converting content from natural languages to UNL (the *EnConverter - EnCo*) and vice versa (the *DeConverter - DeCo*); and
3. System interface components: protocols and tools enabling the flow of UNL documents throughout the web.

The architecture of the system of UNL is represented in figure 1

2.2 UNL Language Components:

The UNL is an artificial language consisting of Universal words (UWs), Relations, Attributes, and UNL Knowledge Base. The Universal words constitute the vocabulary of the UNL, Relations and attribute constitutes the syntax of the UNL and UNL Knowledge Base constitutes the semantics of the UNL. The following subsection will deal with UWs, Relations, Attributes and Knowledge Base respectively. The section will be ended with a concrete example of UNL graph.

2.2.1 Universal Words (UWs): The Vocabulary of UNL

A Universal Word represents simple or compound concepts. UWs are made up of a character string (an English-language word) followed by a list of constraints. There are three kinds of UWs, Basic UWs, Restricted UWs and Extra UWs [3].

2.2.2 Relations: The Syntax of UNL

Binary relations are the building blocks of UNL sentences. They are made up of a relation and two UWs. The relations between UWs in binary relations have different labels according to the different roles they play. The relations are linguistically (semantically) based [4] and similar to those described in [5]. A relation label is represented as strings of 3 characters or less, see the following example:

agt (agent) relation: It indicates a thing in focus that initiates an action. An agent is defined as the relation between UW1 (do) and UW2 (a thing) where UW2 initiates UW1. Consider the following sentence: John breaks the glass. As “John” is the initiator of the action (a thing) and “break” is the event (do), then UW1 will be represented by “break” and UW2 will be represented by “John”. In this case, we can hold an **agt** relation between **UW1** and **UW2**.

2.2.3 Attributes: Expressing Subjectivity of the Speaker

Attributes are mainly used to describe the subjectivity of sentences. They show what is said from the speaker’s point of view: how the speaker perceives what is said. This includes

phenomena technically called “speech acts”, “propositional attitudes”, “truth values”, etc. Attributes are used to describe logicity of UWs, *times with respect to the speaker*, speaker’s view on aspects of event, speaker’s view of reference to concepts, speaker’s view of emphasis, focus and topic, speaker’s attitudes, speaker’s feelings and judgments and attributes for convention.

2.2.4 UNL Knowledge Base: the Semantics of UNL

The UNL KB constitutes the semantic background of the UNL System. It is constituted by the binary direct relations between two UWs. There are three main categories of relations: “icl” (sub class of); “iof” (element/instance of); and “equ” (equivalent to). With these links, a conceptual network can be shaped in the form of UNL thesaurus, UNL ontologies, encyclopedias, or libraries. Its hierarchical structure provides for upper nodes (parent-nodes) and sub-nodes (child-nodes). One single child-node may have several parents-nodes, i.e., the UNL KB allows for a lattice structure.

The hierarchical structure allows for implementing the principle of inheritance in the definition of concepts. All information assigned to a parent-node can be subsumed as inherited by the children-nodes. For instance, in defining “cat(icl>feline)”, there is no need to repeat all properties of felines, as they are defined under all mammals of the same species.

The UNL KB is meant to assure robustness and precision to the UNL System, both to the NL- UNL encoverting, and to the UNL-NL deconverting processes. In the former case, the UNL KB would be used as a sort of word sense disambiguation device. In the latter, the UNL KB would allow for the deconversion of UWs not enclosed in the target language dictionaries. In additional, the UNL KB would enable intelligent search and semantic inference [6]

2.3 Using the UNL in translation

Usual translations are made directly from one language into another. Translation with the UNL system is a two-step process (Inter-lingual).

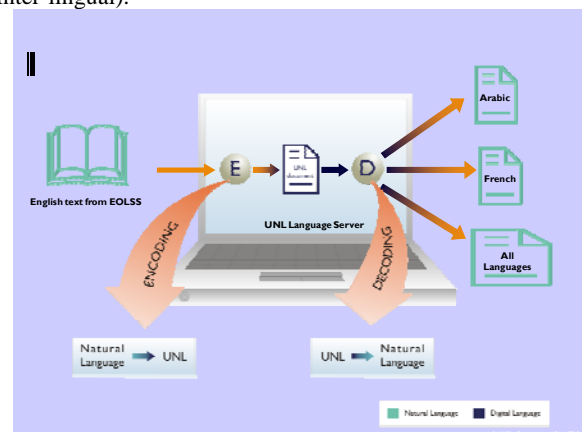


Figure 2: The UNLization process.

The first step deals with Enconverting the content of the EOLSS documents from English (the source language) to UNL (the universal representation). This process is called the *UNLization*

process (figure 2); it is carried out with the use of the English-UNL Enconverter (EnCo). Initially, some post-editing is needed, but as the performance of the English EnCo and the technical dictionaries improve, human intervention will be gradually reduced, and productivity will be increased.

The second step deals with Deconverting EOLSS content from UNL to any natural language according to the native language of the user [7]. This Deconversion process is a task to be carried out by the UNL-Language Server of each language. Each UNL-Language Server contains a dictionary and generation rules (deconversion), working in association with the UNL KB, which are the enabling components in this process.

The system is already in action and could enconvert 25 chapters form EOLSS in UNL. It is a prototype test for translating massive amount of text. The UNL version of EOLSS is sent to the UNL language centers responsible for deconverting them into the six official languages of the United Nations. This work is done in anticipation to the deconversions in many other languages of the world.

The results will be posted in UNESCO web site as a preview of services provided in UNL.

3. Representing documents' metadata (UNL Library Information System)

The proposed module is an art of a multilingual catalog of books metadata. For each language including the UNL, Database (DB) tables are created for the metadata. The system process can be described as follows:

- a) The ingestion phase: The book metadata are ingested for the first time in the metadata DB tables in the language of the book. Starting with the book original language will reduce the possibility of erroneous data.
- b) The encoding phase: The metadata are encoded in UNL representation.
- c) The decoding phase: For each target language, the book metadata are pulled from the UNL DB tables and translated into the DB tables of the target language.

The catalog has, of course, authority tables for authors, publishers, classifications, keywords ...etc. Records of the authority metadata tables are also encoded in UNL. The corresponding new UWs are added to the UNL dictionary [8].

A brief description of the workflow of the proposed system is as follows (LF represents functions utilized by librarians while UF represents functions utilized by UNL specialists):

- The book edition is cataloged in the DB tables in the edition language (LF).
- New entries are added to the Authority tables if needed (LF).
- New UWs are added to the UNL dictionary (UF).
- New metadata records are encoded to UNL and verified if necessary (UF).

The decoding (translation) phase starts by pulling data that should be translated from the UNL catalog and:

- Add new UWs to the target language dictionary, if needed (UF).
- Encode the corresponding new authority file, if needed (UF).

- The metadata are decoded automatically to the target language.
- Verify the translated entries in the catalog as far as the UNL is concerned (UF).
- If the book has an edition in the target language, the metadata of the edition is entered in the system as a new book with a link to the original language (LF).

It is worth mentioning that the following remarks should be taken into considerations:

- Authority DB tables can be built incrementally, but it is advised to start with a reasonably filled authority DB tables.
- The proposed system has the advantage of eliminating duplicate entries in the catalog because data are entered from the original edition and in the same language.
- The authority DB tables must contain different forms of writing author, publisher ... names (e.g. Shakespeare, Shakespeer and Shakspear).
- The proposed system may help in standardizing writing names in a language with different character shapes (e.g. Arabic).
- A good practice is to consult publishers how they would like to write their names in different languages.

As a proof of concept, a prototype system is under development that deals with UNL representation of 1000 books (800 in Arabic, 100 in English and 100 in French). UNL dictionaries needed are already built as well as formal grammars that will encode automatically metadata in terms of universal semantic network. Having finished with this, the encoded metadata will be sent to different UNL language center in order to deconvert it to different natural languages. The prototype system is designed and is under development implemented by Ibrahim Shihata Arabic UNL Language center, International School of Information Science (ISIS), Bibliotheca Alexandrina, Alexandria, Egypt [9].

4. The browsing interface

The other interesting feature of the UNL is its power of searching, storing retrieving, disseminating and sharing data. UNL deals with concepts as metadata which are stored in a "language independent" format. We are developing some special tools for browsing this type of knowledge that offer three modes of navigation: Explore, Discover and Search. For more details about figures 3, 4 and 5, refer to [10].

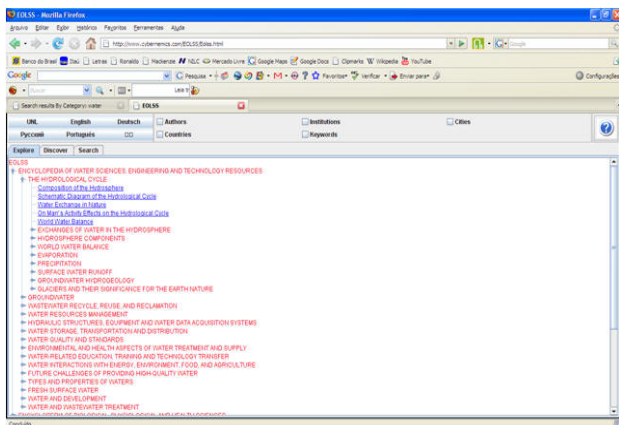


Figure 3: Explore view.

The Explore view (figure 3) shows a tree structure, both expandable and collapsible, representing the Table of Contents. It is the best navigation strategy for those who already know the Encyclopedia internal structure and look for fast and direct access to the EOLSS entries. Entries, when activated, lead directly to the Encyclopedia articles, which can be viewed in the right window when activated

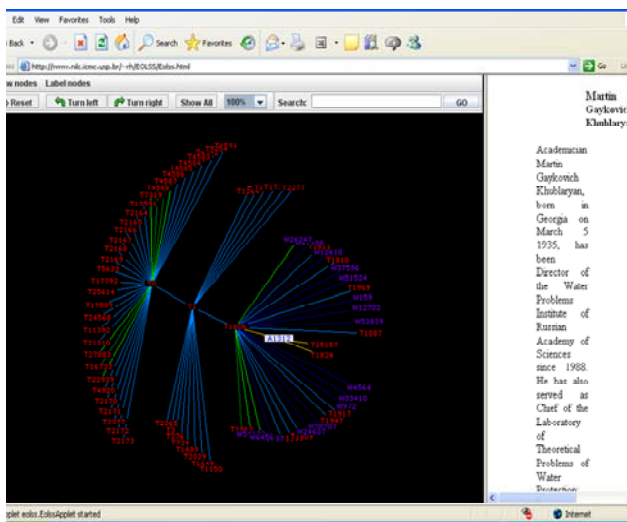


Figure 4: Discover view of EOLSS browser.

In the Discover view (figure 4), we present a hyperbolic tree structure, a navigable graph interlinking titles, keywords, authors, institutions, cities and countries.

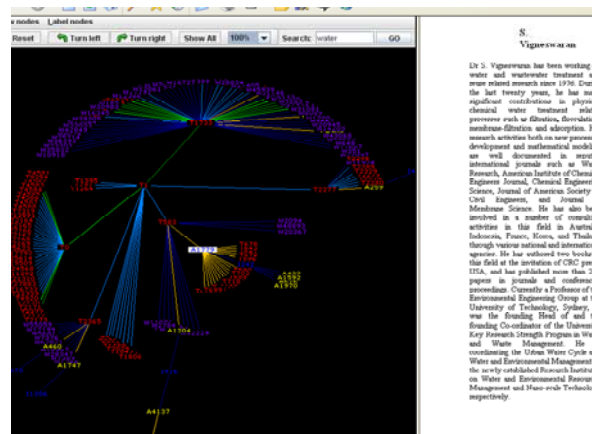


Figure 5: Search view of EOLSS browser.

The Search view (fig 5) performs the search in the EOLSS search engine and displays the results as a new navigable hyperbolic tree, where search can be refined as nodes are again hyperlinks to the articles of EOLSS.

References

- [1] <http://www.eolss.net>
- [2] Uchida, H., Zhu, M., Della Senta, T. (2005), *The Universal Networking Language*. UNDL foundation.
- [3] Uchida, H. (1996), *UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration*. UNU/IAS/UNL Center. Tokyo, Japan.
- [4] Uchida H.(2003). Knowledge Description Language, *Semantic Computing workshop*, Tokyo, Japan.
- [5] Fillmore, C. (1968). The Case for Case. In Bach, E. and Harms, R.T. (orgs.), *Universals in Linguistic Theory*, pp. 1-88. Rinehard and Winston, New York.
- [6] Martins, R. "Knowledge Vertices in XUNL" submitted 2008
- [7] Alansary, S., Nagi, M., Adly, N. (2007), A Semantic Based Approach for Multilingual Translation of Massive Documents, *the seventh International Symposium on Natural Language Processing (SNLP)*, Pattaya, Thailand.
- [8] Alansary, S., Nagi, M., Adly, N. (2006). Towards a Language Independent Universal Digital Library, *2nd International Conference on Universal Digital Library (ICUDL 2006)*, Alexandria, Egypt, 17-19 November, 2006.
- [9] <http://www.bibalex.org/isis/>
- [10] Martins, Ronaldo:
<http://www.ronaldomartins.pro.br/eolss/>