Platform Independent Terminology Interchange Using MARTIF & OLIF

M Ramshirish

Abstract

In the present society, information is treated as one of the most valuable resources and its importance has increased manifold in the networked digital environment. No doubt, the developments in the field of IT have facilitated the sharing of information from one knowledge domain to another, but the difference in terminologies acts as communication barrier and impedes the free flow of information. Various terminological and lexical databases have been developed to help the professionals; however, the problem of data interchange/ exchange across different platforms still persists. This paper gives an overview of MARTIF and OLIF, the tools for platform independent terminology exchange/ interchange.

Keywords : Terminology Database, Localization, Terminology Interchange, MARTIF, OLIF

0. Introduction

Wide spread growth of human knowledge and the growth in the field of Information Technology has led to the preparation of many databases of terms and lexicons. Development of term databases and lexicons is time consuming and a costly affair. Moreover, differences in software, hardware, and methodology further complicate the interchange of common terms across databases. To find a solution to these problems and to make the web of chaos “a semantic web”, the idea of exchange/interchange of terminology and lexicons came up.

Terminology is a set of terms used by the subject specialists and experts in their respective specialized areas. It provides base from which original technical texts and translations are prepared. Technical terminology is useful not only for translators and technical writers as document producers, but also for others like teachers who must be able to guide their students in acquiring the specialist knowledge.

MARTIF (Machine Readable Terminology Interchange Format) and OLIF (Open Lexicon Interchange Format) have been developed to solve the problems of platform-independent terminology interchange between different databases.

1. What is Terminology

In any subject, the subject specialists use technical terms and other terms that are related to that particular subject, these specialized terms constitute the terminology.

Terminology is involved in one or the other way, whenever and wherever, specialized information and knowledge is created, communicated, recorded, processed, stored, transformed, re-used.

Terminology can be defined as “a structured set of concepts and their relations in a particular subject field. It can also be considered as the infrastructure of specialized knowledge”. [1]

1.1 Why Terminology

Due to increase in the R&D activities, rapid growth of Internet, literature, general people as well as researchers have seamless access to a variety of information from different sources. But, if people fail to understand the information and data, then this may lead to chaos.

To solve this problem, need was felt to develop terminologies related to a particular subject area where the terms will help one in understanding the concept and relations between different facets of a subject.

1.2 Terminology Databases and their Significance

In any subject field, new terms are created and also some terms become obsolete. The conceptual meaning of terms also may change with time and developments that take place in a subject field. Some terms have very precise meaning in some situation, and after many years they are used in a different meaning associated with old usage [2]. E.g.: - The term Ontology was used earlier in philosophy to describe the nature of existence, but now it means construction of knowledge models with specific concepts or objects, their attributes, and inter-relationships.

So, terminology databases were developed to overcome the above problems; the databases attempted to rationalize the use of terms in technical language. In them, confusing terms such as homonyms were avoided to facilitate exclusive usage of terms in a very particular subject area and to achieve vocabulary control as a result.

It is impossible to think of technical writing and technical documentation without the knowledge of proper terminology. Terminologies used in a specialized subject area contain not only terms but also formulae, symbols, even drawings where communicating potential of the document lies in the terminology used. Adequate use of terminological resources increases quality and productivity for documentation. Further, we need suitable terms to get exact meaning from the original document for translation of any document from one language to another.

1.3 Users of Terminology Databases

Some standard format is needed to ensure reliable and qualitative exchange and for facilitating interoperability. Different groups of professionals use different terminology databases for various purposes.

They can be divided to following broad categories [1]: -

Translators

Technical writers

Information managers

Professionals dealing with controlled vocabularies tools like thesauri, classification Systems, etc. for documentation and information retrieval

They use the terminology databases for [3]: -

- ✍ Data capture and presentation – i.e., enter, store and review the concepts. Information integration, indexing, retrieval, decision support, linking records. Messaging between software systems i.e., linking different information systems.
- ✍ Reporting.

1.4 Need of Terminology Interchange [4]

In order to exchange expert information and prevent duplication of efforts, users of terminological databases need interchange formats, as preparation of terminological database is both time and money consuming.

The terminology databases are implemented in different formats and they are prepared to run on different operating systems, requiring sometimes-different file format and sign conventions. So, a standard format is needed for terminology interchange for preparation of data interchange tools.

A common format has to be defined for terminology export and import to provide a channel for terminology interchange. Possibly terminology interchange could be managed without any structure, i.e. an unstructured text file, but this leads to the problem of reformatting the unstructured data manually; a time and cost intensive undertaking.

2. Need of Standards for Terminology and Lexicon Interchange

The exchange of database records needs to be done carefully because structure of terminological records varies considerably from one database to another. In addition to this, even the designs and user needs vary, so, here a universal interchange format is essential to make interchange easier.

But the language-processing tools are not well integrated and interoperable. Terminology databases, translation memory systems, controlled English systems, machine translation systems lack seamless integration. The desired integration, once achieved, would increase productivity among translators, terminologists and other workers. So, the tools for terminology extraction, terminology consistency checking and translation workflow are needed. [5]

To solve these problems, various standard formats (including MARTIF and OLIF) evolved for modeling and representing the terminological data.

3. MARTIF & OLIF

3.1 What is MARTIF

MARTIF stands for Machine Readable Terminology Interchange Format. It is a SGML based format to facilitate the interchange of terminological data among Terminology Management systems [4].

Its origin can be traced back to the development of a powerful tool for terminology interchange, done in cooperation with Text Encoding Initiative (TEI) and Localisation Industry Standards Association (LISA) [LISA is an association of companies and institutions working on the translation and adaptation of software into different languages and TEI is an initiative for encoding texts with a relevant structure and semantic information]. The goal of this cooperation was to produce a format that would be a platform independent and publicly available format. The resulting format was MARTIF, which is also known as ISO (FDIS i.e., Final Draft International Standard) 12200 (which is in turn based on ISO 12620 [ISO 12620 is designed to promote consistency in the storage and interchange of terminological data through the use of a standard set of data categories for term entries]).

3.2 Evolution of MARTIF [6]

Modern developments in Terminology exchange field can be seen from an earlier format called MATER (ISO 6156) (Magnetic Tape Exchange for Terminological lexicographic Records (1987), which was published only after magnetic tape become obsolete. MATER developed to MicroMATER (designed to meet the needs of the then new PC architectures). The various versions, which came up in due course are MARTIF part 1 (Negotiated Interchange) and MARTIF part 2 (Blind Interchange).

In the first approach, two partners use a common framework for interchange and negotiate details within the framework of the intermediate format to allow writing export and import routines that preserve as much information as possible. Presently work is going on in the area of second approach i.e. MARTIF part 2 (1998) (Blind Interchange). In this, the details are predefined, so, export and import routines can be written without knowing who the other interchange partners will be (i.e., one need not 'see' the interchange partners).

3.3 Purpose of MARTIF [7]

To provide a universally applicable format for the negotiated interchange of structured terminological data among various applications, system environments, and hardware platforms.

To be used with terminological data that can be stored, read, retrieved, and manipulated by a computer (ISO DIS 12200.2:1).

3.4 Features of MARTIF [8]

MARTIF is concept-based approach rather than non-concept oriented approaches to terminology e.g., lexicographic and NLP approaches.

It allows directional change in using the database or in importing the data and readily allows for the interchange of data to or from other data models where different languages are given preference.

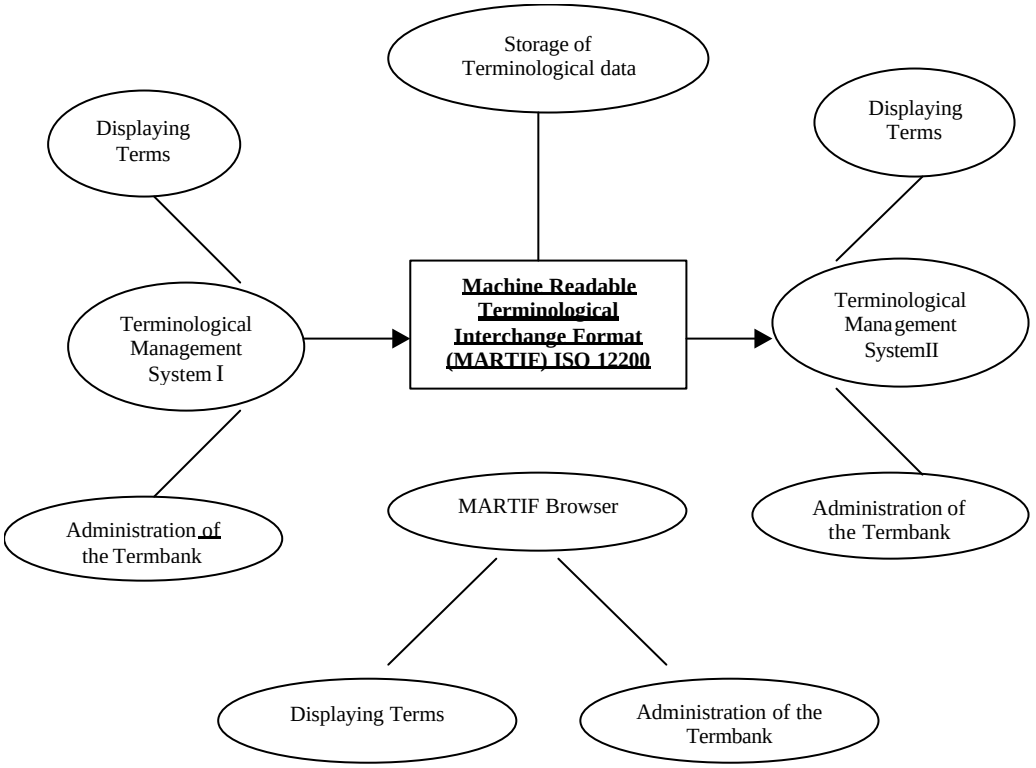
3.5 MARTIF Specifications

Categories for MARTIF are divided into ten sections, which are grouped into four classes. The four classes with their sections are [4]:

1. Term: Section(1)- consists of the data category term;
2. Term-related information: Section (2)- consists of term-related information (such as POS, etymology, term type, ...) and section (3)- on the degree of equivalence (how close two terms are related);
3. Descriptive information: Section (4)- relation to the domain, section (5)- descriptions of the concept (i.e. definitions, examples), section (6)- relations between concept entries, section (7)- data categories that relate a concept entry to its position in the concept system, section (8)- general notes;
4. Administrative information: section (9)- consists of data categories relating a concept entry to a node in a thesaurus or to other forms of documentation, section (10)- has data categories which contain administrative information (update information, author, etc.).

3.6 Usage of MARTIF

The usage of MARTIF can be depicted in following diagrammatic manner



4. What is OLIF?

It can be defined as Open Lexicon Interchange Format, it is an XML compliant standard [9].

A lexicon can be defined as

1. The collection of words in a language
2. It is a special list of terms and related terms used for subject searches.
3. It is a linguistic tool with information about the morphological variations and grammatical usage of words.

4.1 Puspose of OLIF

OLIF allows the transfer of terminological and lexical data between or from different translation tools, including NLP systems such as Machine Translation as well [10].

4.2 Origin of OLIF

OLIF has its origin in the Open Translation Environment for Localization (OTELO) project, which was funded by European commission sharing lexical resources [10]. Members of OLIF Consortium designed it and are the maintaining organization.

The OLIF consortium and SALT group (Standard based Access Service to multilingual Lexicons and Terminologies) are collaborating with each other. Here, SALT aims to create a lexicon exchange format and SALT focuses on the terminological side of the format (in the tradition of MARTIF), while OLIF focuses on the lexical side.

OLIF and SALT's XML based formats for Lexicon and Terminologies (XLT) define a common set of data categories to enable integration between OLIF and XLT. It is an organization of major NLP technology suppliers, corporate users of NLP and research institutions, Systran, Logos, Sail Labs, IBM/Lotus, Lingua Tec, Pa Trans, Trados, Xerox, German Research Centre for Artificial Intelligence, IAI and Others etc.

4.3 Features of OLIF [10]

OLIF is more practical and tries to accommodate existing lexical resources while other lexicon projects are more research oriented, focusing on elaborated lexical descriptions.

OLIF concentrates on Lexical exchange rather than terminology, and leans more towards pragmatic than theoretical or research based projects.

The basic idea of OLIF is to facilitate the exchange of primarily the pivotal information in lexical entries. It also provides the option of deeper lexical representation included in the OLIF format

It offers the user a mechanism for encoding the information in a general way that allows portability.

4.4 OLIF Specifications

The basic unit in OLIF is uniquely defined by a set of key data categories: Canonical form, parts of-speech, language code, subject area, and, in the case of homonyms a semantic reading etc [11].

5. Conclusion

MARTIF & OLIF are not the only formats which facilitate the terminological and Lexical Interchange, other formats also exist, but they are not powerful enough to handle the complexity of the problem. MARTIF uses SGML structures in order to preserve and communicate the information and functionalities present in many terminological databases whereas OLIF uses XML. The standards are no doubt complex for the end users to understand but the complexity is required to facilitate terminology exchange among different databases.

6. References

1. Galinski, Christian & Budin, Gerhard.Terminology. http://cslu.cse.ogi.edu/HLTsurvey/ch12node7.html (accessed on 15/1/04)
2. Electronic Dictionaries DicoBase. http://www.linga.fr/LingEn/DicoOffreEn.htm. (accessed on 25/10/2004)

3. Rector, Alan. Why do we need Medical Terminologies? http://www.cs.man.ac.uk/mig/links/RCSEd/terminology-why.htm (accessed on 09/12/2003)

4. Trippel, Thorsten. Terminology interchange: Facing multiple requirements. http://coral.lili.uni-bielefeld.de/~ttrippel/terminology/node82.html#SECTION008400000000000000000 (accessed on 26/09/2004)

5. Warburton, Kara.Results of the LISA Terminology survey. http://www.lisa.org/2001/termsurveyresults.html (accessed on 08/11/2004)

6. Robin Bonthrone, Fry & Partnerschaft, Bonthrone. MARTIF Lite: User-driven Terminology Interchange. http://www.lisa.org/archive_domain/newsletters/1998/1/bonthrone.html (accessed on 30/09/2004)

7. The CLS Framework:Negotiated Sharing. Introduction to ISO 12200 (negotiated MARTIF) http://www.ttt.org/clsframe/negotiated.html (accessed on 26/10/2004)

8. MARTIF putting complexity in perspective. http://www.ttt.org/clsframe/termnet1.html (accessed on 30/09/2004)

9. www.olif.net (accessed on 19/10/2004)

10. Lieske, Christian.Cormick, Susan Mc & Thurmair, Gregor. The Open Lexicon Interchange Format (OLIF) comes of Age. http://www.eamt.org/summitVIII/papers/lieske.pdf (accessed on 29/09/2004)

11. OLIF. http://www.w3.org/2002/02/01-i18n-workshop/OLIFExample.xml (accessed on 08/11/2004)

12. MARTIF. http://www.creativyst.com/cgi-bin/M/Glos/st/GetTerm.pl?fsGetTerm=802.1b (accessed on 07/10/2004)

13. Terminology Interchanges. http://korterm.org/research3-e.htm (accessed on 07/10/2004)

About Author

Mr. M. Ramshirish is working as Librarian at RRG E-Media, Ramoji Film City, Hyderabad. He has done ADIS from DRTC, Indian Statistical Institute, Bangalore and BLISc. from Osmania University, Hyderabad.

E-mail : ramshirish@yahoo.co.in or mramshirish@yahoo.co.in