MARTIF: New ISO Standard for Translation

By
Klaus-Dirk Schmitz and
Sue Ellen Wright

A proposed standard for terminology exchange to address a growing need.

Target-oriented, efficient specialized communication requires correct terminology. Consequently, subject-area experts, technical communicators, and documentation and information specialists need access to specialized dictionaries (usually monolingual ones) that provide domain-specific definitions and explanations

When special language communication takes place across linguistic boundaries, translators and interpreters must translate texts into target languages in a way that meets the assumed needs of the target audience. Research in multilingual terminology resources is a prerequisite for high-quality translation products. Language planners, standardizers, special-language lexicographers, and terminologists support terminology consumers by collecting, processing, and documenting mono- and multilingual technical vocabularies.

The traditional media for collecting, disseminating, and using terminologies, such as specialized dictionaries, glossaries, and card file collections, are rapidly yielding ground to computer database-management solutions. This trend began in the 1960s with the creation of mainframe-based terminology databases and continues today in the form of numerous, mostly PC-based, terminology database-management programs designed to meet the full range of user needs.

The creation of high-quality terminology products is both time-consuming and cost-intensive. Consequently, the community of terminology users has a vested interest in exchanging terminology data collections. However, different user groups and organizational environments have different requirements so far as choice of languages and information categories is concerned, with the result that any two terminology collections are likely to exhibit very different structures, which significantly impairs the feasibility for interchange among different users and systems. Up to now, it has been necessary to write individual conversion programs to accommodate each individual set of partners desiring to exchange data.

Both national standards institutes and the International Standards Organization (ISO) recognized this problem and defined the exchange of terminological data at the beginning of the 1980s (ISO 6156: Magnetic tape exchange format for terminological/lexicographical records (MATER); DIN 2341-1). This standard is, however, not well suited to the exchange of data among modern terminology-management systems, primarily, but not exclusively, because it focuses on the now outmoded handling of data stored on magnetic tape. With a few exceptions, e.g., the exchange of data between LEXIS, Siemens AG's terminology data bank, and EURODICAUTOM, the terminology data bank of the European Union, MATER, has not been used in practice as an exchange standard.

In an effort to address the need for a state-of-the-art standard, ISO is currently nearing the completion on a new exchange format called MARTIF (ISO 12200: Terminology—Computer applications—Machine-readable Terminology Interchange Format, or MARTIF). MARTIF is based on Standard General Markup Language (ISO 8879, SGML) and was originally developed in close cooperation with the Text Encoding Initiative (TEI) and the Localization Industry Standards Association (LISA).

The main body of the MARTIF standard specifies the formalism to be used in preparing terminology data collections for interchange by defining the SGML Document Type Definition (DTD) and listing the appropriate tags (markup) used to structure the data. Normative Annex A of the standard specifies the markup for the individual terminological data categories to be used in the MARTIF environment. Annex A is based on ISO 12620 (Terminology—Computer applications—Data categories), which has been developed parallel to the MARTIF standard to define the data categories used in terminology collections.

The illustration shows an example of a MARTIF document. This text is somewhat difficult to read at first glance because it uses "character entities" to represent accented characters that are not represented in standard ASCII. For instance, the German phrase "Maß für die Lichtundurchlässigkeit" becomes "Maß für die Lichtundurchl&aumlt;ssigkeit" and French

(continued on page 23)

MARTIF In Action: Sample Code

```
<!DOCTYPE martif PUBLIC "ISO 12200:1997//DTD for MARTIF (framework) //EN" [
<!ENTITY % mtf-body.PUBLIC "ISO 12200:1997//DTD for MARTIF (base) //EN" >
<!ENTITY % mtf-ents PUBLIC "ISO 12200:1997//ENTITIES for MARTIF (sets) //EN" >
]>
    <martif>
        <martifHeader>
            <fileDesc>
                <titleStmt><title>Example 1: a complete martif document</title></titleStmt>
            </fileDesc>
        </martifHeader>
        <text>
            <body>
                <termEntry>
                        <descrip type='subjectField'>appearance of materials</descrip>
                        <ntig lang=en>
                                <termGrp>
                                        <term>opacity</term>
                                        <termNote type='partOfSpeech'>n</termNote>
                                        <termNote type='termType'>preferred term</termNote>
                                </termGrp>
                        </ntig>
                        <ntig lang=de>
                                <termGrp>
                                        <term>Opazität</term> <termNote type='partOfSpeech'>n</termNote>
                                        <termNote type='gen'>f</termNote>
                                </termGrp>
                                <descripGrp>
                                        <descrip type='definition'>Maß für Lichtundurchlässigkeit</descrip>
                                        <ref type='sourceIdentifier' target='DIN6730.1992-08'>p. 5</ref>
                                </descripGrp>
                        </ntig>
                </termEntry>
            </body>
            <back>
                <bibl id='DIN6730.1992-08'>Papier und Pappe: Begriffe</bibl>
            </back>
        </text>
    </martif>
```

MARTIF (continued from page 18)

"contrôle" becomes <contrôle". Of course, human readers will find these conventions confusing, but format designers intend that the first step in using imported data will be for local users to convert these codes into recognizable characters. Some of the features of this sample reflect the latest comments received from the MARTIF working group.

MARTIF provides an open, flexible format for the exchange of terminological data among different terminology database-management systems. MARTIF can be used for more than just the exchange of data between different users—it can be employed when companies need to change or upgrade software from one database format to another. MARTIF's SGML base also makes it easier to transfer data to other SGML documents using the new interchange standard, for instance for the publication of dictionaries. Furthermore, the SGML base of MARTIF documents provides an excellent springboard for transferring terminological data to HTML environments for "publication" on the World Wide Web, a process that is currently being tested by the Virtual Hypertext Glossary project in the UK.

ISO 12200 and ISO 12620 go a long way toward defining terminological data for interchange purposes, but, aside from a limited number of restrictions, such as standardized data formats and language codes, the guidelines set down in these standards do not attempt to deal with the content of data categories. Consequently, it can be assumed that additional information about the content of the data categories will have to be provided for the terminology database-management systems involved in any instance of data interchange if interchange partners are going to be able to avoid information loss or adulteration. It does not look as if even the MARTIF standard will facilitate so-called "blind interchange" between all systems, i.e., the exchange of data between systems whose content and data-modeling characteristics are totally unknown to the exchange partners.

In an effort to test the robustness of the standard and to explore its limitations, various research groups have begun a series of empirical tests using the MARTIF standard for the exchange of terminological data among existing terminology collections based on database-management systems with different entry structures and data categories. One of the goals of this effort is to determine whether "blind interchange" without additional conversion routines is possible or whether additional specifications must be made beyond the MARTIF Standard in order to facilitate terminology interchange.