TERMINOLOGY INTERCHANGE USING MARTIF

Klaus-Dirk SCHMITZ

Fachhochschule Cologne, Germany

1. THE NEEDS FOR TERMINOLOGY INTERCHANGE

Efficient target-oriented specialized communication is inconceivable without 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 convert texts into target languages in a way that meets the assumed needs of the target audience. Research in multilingual terminology resources are 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 is both time-consuming and cost-intensive. As a consequence, the community of terminology users has a vested interest in exchanging terminological data collections. Different user-group needs and organizational environments dictate, however, that the languages and information categories required by individual systems will vary considerably, which means that the structure of different terminology databases will also exhibit a great deal of diversity. As a result, any exchange of terminological data between different systems becomes significantly more difficult than one might anticipate. In the past, these problems have made it necessary for exchange partners to create individual conversion programs to accommodate each exchange situation.

2. THE MARTIF CONCEPT

Both national standards institutes and the International Organization for Standardization (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 (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 Localisation Industry Standards Association (LISA) (see Budin/Melby/Wright 1993).

```
<!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'>noun</termNote>
                        <termNote type='termType'>preferred term</termNote>
                    </termGrp>
                </ntig>
                <ntig lang=de>
                    <termGrp>
                        <term>Opazität</term>
                        <termNote type='partOfSpeech'>noun</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>
                <ntig lang=fr>
                    <termGrp>
                        <term>opacité</term>
                        <termNote type='partOfSpeech'>noun</termNote>
                        <termNote type='gen'>f</termNote>
                    </termGrp>
                </ntig>
            </termEntry>
        </body>
        <back>
            <bibl id='DIN6730.1992-08'>Papier und Pappe: Begriffe</bibl>
        </back>
    </text>
</martif>
```

Figure 1: *Sample MARTIF Document*

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.

Figure 1 shows an example of a MARTIF terminology interchange document. For the sake of readability, this example has retained standard German and French diacritics rather than adopting the SGML entities typically used in MARTIF, i.e., “Tür” for “Tür” or “contrôle” for “contrôle”.

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 spring board 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 (West/Murray-Rust 1996).

3. TESTING MARTIF AND FUTURE DEVELOPMENTS

During the long process of defining and testing the current version of the standard, there has been a serious discussion within the MARTIF work group itself concerning differing fundamental philosophies with respect to interchange, specifically regarding the question of the normative rigor of the standard. MARTIF was originally designed for so-called *negotiated interchange*, where partners examine each other’s data before interchange and make decisions about preconditioning the data before importing it from the interchange format. This approach reflects the notion that an interchange format should be able to accommodate all or almost all existing database structures without imposing standardization on specific systems.

This approach allows for a high degree of freedom and flexibility in individual applications, which some standardizers find disturbing. They would rather displace the effort to the other end of the spectrum, imposing a higher degree of normalization with respect either to the source database structures or at least to the export product from these systems. In response to these concerns, researchers in the US development team have developed and successfully tested a full prototype for a *blind interchange* DTD based on the existing DTD for negotiated interchange. Testing involved developing a single routine for converting data to and from DANTERM, TERMIUM, and MultiTerm (Hardman 1996; Melby/Hardman 1996), and between TermStar and MultiTerm (Reinke/Schmitz 1997).

This second strategy for interchange would enable exchange partners who had agreed to conform to certain specific criteria to import each other's data without prior examination. Such a scenario would require that exchange partners subscribe to a stricter entry layout and, in the case of some data categories such as *subject field*, *gender*, etc., to adopt harmonized content options. The blind interchange option may even impose the requirement that database operators change their local data structures and content in order to participate in interchange. Concerns related to the development of a blind standard have been treated in other contexts and will not be the focus of this article (Wright 1996; Hardman 1996).

Based on the experience derived from testing the MARTIF standard as defined now, ISO TC37/SC3 has decided on its last meeting in Copenhagen to work on a Part 2 of the standard. For this Part 2 certain imprecisions in the data category set would have to be resolved, specifically with regard to granularity, modeling variance, and the content of permissible instances.

Granularity indicates the degree of fine detail included in a database for a given data category. For example, in the case of abbreviated form of term, the current standard offers two options for reporting information, either to use the broader category <termNote type='termType'>abbreviated form</termNote> or to select the more specific permissible instances, i.e., *abbreviation*, *initialism*, *acronym*, *short form*, etc. For Part 2, one of these options would have to be specified as required.

The current standard also provides for variation in *data modeling practice*. For instance, in the case of *variant*, there is the possibility to use either <ptr type= 'variant' target='ID1234'> or <termNote type='termType'>variant. In part 2 of the standard, only the latter option would be allowed.

Here too, precise content of data categories not defined in ISO standards 12620 and 12200 would be specified. For instance, in the case of "grammatical gender", it is not enough to say that "masculine, feminine, neuter, other" are the acceptable values, but rather that specific forms of these values are required, for example: *m, f, n, o*. In Part 2, no attempt would be made to normalize subject field, thesaurus, and classification systems.

For Part 2 of the MARTIF standard, a new DTD has to be developed. This new DTD is planned to be fully compatible to the "negotiated" DTD defined in Part 1 of the standard. Compatibility in this context means that all MARTIF interchange documents conforming to Part 2 of the standard could also be parsed by the Part 1 DTD, but not vice versa. In addition to the DTD, a validation tool has to be developed to control the permissible instances of certain data categories which could not be done by a DTD.

Within the framework of the two parts of the standard, or even beyond the requirements stated in the two parts, individual user groups can agree among themselves to accept additional standard conventions. For instance, they might agree to use a specific subset of the data categories listed in ISO 12620. Models for cohesive user groups (VHG, LISA, etc.) can adopt very similar, but simplified structures while maintaining a back-door link to MARTIF. Such solutions may utilize a simplified tag-naming convention that nevertheless remains parallel to MARTIF in order to facilitate easy conversion. HTML implementations are likely to fit into this pattern, with fundamental variations designed, for instance, for either read-only or interactive use or even the creation of data resources on intranets or via the World Wide Web.

REFERENCES

Budin, Gerhard, Melby Alan, and Wright, Sue Ellen (1993): "Terminology Interchange Format (TIF): A Tutorial." *TermNet News.* No. 40/193, 5-63.

Hardman, Daniel (1996): A Practical Proposal for the Blind Interchange of Terminological Data. Unpublished Master's Thesis. Provo, Utah: Brigham Young University.

ISO DIS 12 200.2 (1995): Terminology—Computer Applications—Machine-readable Terminology Interchange Format (MARTIF). Geneva: ISO TC 37/SC 3/WG 3.

ISO DIS 12620 (1995): Terminology — Computer Applications — Data Categories. Geneva: ISO TC 37/SC 3/WG 1.

Melby, Alan, and Hardman, Daniel (1996): "Importing Terminology from Multiple Sources in Three Phases: Inspection, Adjustment and Adoption." *Proceedings of TKE '96: Terminology and Knowledge Engineering.* Christian Galinski and Klaus-Dirk Schmitz, Editors. Frankfurt/M.: Indeks Verlag, 197-204.

Melby, Alan, Schmitz, Klaus-Dirk, and Wright, Sue Ellen (1996): "The Machine Readable Terminology Interchange Format (MARTIF)—Putting Complexity in Perspective." *TermNet News*. No. 54/55/1996, 11-21.

Reinke, Uwe, and Schmitz, Klaus-Dirk (1997): "Testing the Machine Readable Terminology Interchange Format (MARTIF)". *Saarbrücker Studien zur Sprachdatenverarbeitung.* Saarbrücken: Universität des Saarlandes.

Schmitz, Klaus-Dirk (1996): "Martif: A New ISO—Standard for the Interchange of Terminological Data." Translated by Sue Ellen Wright. *TermNet News*. No. 50/51/1995, 6-8.

Schmitz, Klaus-Dirk, and Wright, Sue Ellen (1997): "Terminology Interchange - Needs, Approaches, Solutions, Problems". *Proceedings of the 2nd International Conference on Terminology, Standardization and Technology Transfer TSTT'97.* Beijing: Encyclopedia of China Publishing House, 346-353.

West, Lesley J., and Murray-Rust, Peter (1996): "Terminology in a Global Context—The Virtual Hypertext Glossary (VHG)". *TermNet News*. No. 50/51/1995, 6-8.

Wright, Sue Ellen, and Budin, Gerhard (1994): "Data Elements in Terminological Entries: An Empirical Approach". *Terminology*. Vol. 1, No. 1, pp. 41-60.

Wright, Sue Ellen (1996): "Blind Interchange of Terminological Data: Problems and Possibilities," *Multilingualism in Specialist Communication: Proceedings of the 10th LSP-Symposium*, 29 August - 2 September, 1995. Vienna: TermNet., 1123-1130.

Wright, Sue Ellen (1997): "Mapping Local Data Categories to Categories Defined in ISO 12620." *IITF Journal*, in press.

****

LABURPENA / RESUMEN / RÉSUMÉ / ABSTRACT

Terminologi trukea MARTIF formatuaren bidez

Kalitate oneko terminologia sortzeak denbora eta diru asko eskatzen ditu. Horren ondorioz, terminologia erabiltzen dutenek interes pertsonala daukate datu terminologikoen multzoak trukatzean. Hala ere, erabiltzaile-multzo desberdinen beharrak eta antolaketaren baldintzak direla-eta, banakako sistemek behar dituzten hizkuntzak eta informazio-kategoriak asko aldatzen dira. Hortaz, datu-base terminologikoen egiturek ere desberdintasunak erakusten dituzte. Horren ondorioz, sistema desberdinen artean egiten den datu terminologikoen trukea aurreikusi baino askoz zailagoa da. Iraganean, arazo horiek zirela-eta, trukean parte hartzen zutenek banakako bihurketa-programak sortu behar izaten zituzten trukaketa bakoitza egokitzeko.

Kontzeptuekin lotutako sarrera terminologikoak dauzkaten datu-multzoak trukatzea unibertsalagoa izan dadin, eta ez hain garestia, ISO 12200 arauan (Final Draft International Standard izenekoan) MARTIF formatua sortu da (Machine-Readable Terminology Interchange Format). Formatuaren oinarria, hein handi batean, kide den ISO 12620 (FDIS) arauko datu-kategorien izen eta definizioetan dago. Azken arau horrek, funtsean, datuen eremu batzuetan eduki moduan sar daitezkeen datuen eremuen izenak eta adibide zilegiak zehazten ditu. MARTIF formatua Standard Generalized Markup Language hizkuntzan oinarritzen da (ISO 8879, SGML).

Ponentzia honek terminologiaren trukea egiteko dauden arrazoiak emango ditu eta truke-formaturako eskakizun zehatzak aurkeztuko ditu. MARTIF eta datuen kategoriari buruzko arau egokia adibideen bidez azalduko dira, eta datu-base terminologikoetan datuei forma emateko modu desberdinek dakartzaten arazoak aztertuko dira. Aurkezpenaren

osagarri modura, etorkizunean truke-maila desberdinak garatzeko aholku batzuk emango dira.

Intercambio de terminología por medio de MARTIF

La creación de terminología de alta calidad exige mucho tiempo y dinero. Como consecuencia, la comunidad de usuarios de terminología tiene un interés personal en intercambiar conjuntos de datos terminológicos. Sin embargo, las necesidades de los diferentes grupos de usuarios y los entornos organizativos imponen que las lenguas y categorías de información que necesitan los sistemas individuales varíen considerablemente, lo cual significa que la estructura de las diversas bases de datos terminológicas también mostrarán gran diversidad. Como resultado, cualquier intercambio de datos terminológicos entre sistemas diferentes se hace mucho más difícil de lo que se puede prever. En el pasado, esos problemas obligaron a los socios en el intercambio a crear programas de conversión individual para adaptar cada situación de intercambio.

Para lograr que el intercambio de conjuntos de datos que contengan entradas terminológicas dirigidas a conceptos sea más universal y menos costoso, se ha elaborado en la norma ISO 12200 (Final Draft International Standard-FDIS) el Machine-Readable Terminology Interchange Format (MARTIF). El formato se basa en gran medida en los nombres y definiciones de las categorías de datos contenidos en la norma afín ISO 12620 (FDIS), que en esencia especifica los nombres de campo de los datos y ejemplos lícitos relacionados que pueden incluirse como contenidos en algunos campos de datos. MARTIF se basa en el formato SGML (Standard Generalized Markup Language-ISO 8879).

La ponencia argumenta por qué hay una necesidad general para un intercambio de terminología y cuáles son las demandas concretas para un formato de intercambio. MARTIF y la correspondiente norma de categoría de datos se explicarán e ilustrarán por medio de ejemplos, y se debatirán los problemas sobre diversas formas de moldear los datos en las bases de datos terminológicas. La presentación se completará con recomendaciones para el futuro desarrollo de diferentes niveles de intercambio.

Échange de terminologie utilisant MARTIF

La création d'une terminologie de grande qualité non seulement prend du temps mais également est coûteuse. Par conséquent, la communauté des utilisateurs de terminologie est légitimement fondée à échanger des collections de données terminologiques. Différents besoins tenant aux groupes d'utilisateurs et à leurs environnements d'organisation font, toutefois, que les langages et catégories d'information requise par des systèmes individuels varieront considérablement, ce qui signifie que la structure de différentes bases de données terminologiques montrera également une bonne part de diversité. Le résultat en est que tout échange de donnée terminologique entre différents systèmes devient significativement plus difficile qu'on ne pourrait l'imaginer en première instance. Dans le passé, ces problèmes ont rendu nécessaire pour les partenaires voulant échanger de créer leurs propres programmes personnels de conversion afin de pouvoir s'adapter à chaque situation d'échange.

Dans le souci de faciliter un échange plus universel et moins coûteux de collections de données contenant des entrées de terminologie orientées vers les concepts, le format MARTIF (Machine-Readable Terminology Interchange Format) a été élaborée en ISO FDIS 12200 (Final Draft International Standard). Le format repose fondamentalement sur les noms de catégories de données et définitions contenus dans le standard associé ISO FDIS 12620, lequel spécifie essentiellement des noms de données de champs et en rapport avec des exemples admissibles susceptibles d'être inclus comme contenus de certains

champs de données. MARTIF est basé sur le Standard Generalized Markup Language (ISO 8879, SGML).

L'article plaide les raisons pour lesquelles on se trouve face à un besoin général en matière d'échange de terminologie en expliquant en quoi les demandes spécifiques exigent un format d'échange. MARTIF et le standard de catégorie de données correspondantes seront expliqués et illustrés par des exemples, et les problèmes donnant lieu à différentes variantes de modèles de données en matière de bases de données terminologiques seront discutés. Des recommandations pour un développement dans le futur de différents niveaux d'échange compléteront la présentation.

Terminology interchange using MARTIF

The creation of high-quality terminology is both time-consuming and cost-intensive. As a consequence, the community of terminology users has a vested interest in exchanging terminological data collections. Different user-group needs and organizational environments dictate, however, that the languages and information categories required by individual systems will vary considerably, which means that the structure of different terminology databases will also exhibit a great deal of diversity. As a result, any exchange of terminological data between different systems becomes significantly more difficult than one might anticipate. In the past, these problems have made it necessary for exchange partners to create individual conversion programs to accommodate each exchange situation.

In order to facilitate more universal, less costly exchange of data collections containing concept-oriented terminology entries, the Machine-Readable Terminology Interchange Format (MARTIF) has been elaborated in ISO Final Draft International Standard (FDIS) 12200. The format relies heavily on the data category names and definitions contained in the companion standard ISO FDIS 12620, which essentially specifies data field names and related permissible instances that may be included as the contents of some data fields. MARTIF is based on Standard Generalized Markup Language (ISO 8879, SGML).

The paper argues why there is a general need for terminology interchange and what the specific demands are for an interchange format. MARTIF and the corresponding data category standard will be explained and illustrated by examples, and problems involving different data modeling variances in terminological data bases will be discussed. Recommendations for future development of different levels of interchange will complete the presentation.