Gregor Thurmair

Exchange Formats: TBX, OLIF, and Beyond

Abstract
This paper tries to comment on some of the standardisation efforts in the area of exchange formats for lexical resources. The first family of standards was centred around terminological data, producing exchange formats like MATER/MARTIF and TBX, based on an organisation of the data as concepts and (language-specific) terms. When the exchange of fully annotated lexical data came into play, standards like OLIF and MILE were proposed; they focus on the representation and the exchange of (mono- and multilingual) dictionary entries and their attributes (Thurmair/Lieske 2002). Recent developments are organised around the creation of markup frameworks, try to define frameworks for meta-models on one hand, and sets of elementary data categories on the other hand, both of which can be grouped into workable exchange formats.

1 TBX

1.1 History

The first exchange format for terminology was called MATER; it defined how data had to be stored on a magnetic tape, specifying, among other things, byte sequence, tape length, block size etc. This format was converted into Micro-Mater (for PC exchange), and later into MARTIF, the first SGML-based format. Martif underwent several standardisation steps (ISO 12200, ISO 12620 and others) and was further developed in an EU funded project called SALT. The current status of the format is XLT (XML-based Formats for Lexicon and Terminology Exchange) which is the framework for several flavours of the standard depending on the different use cases; the most widely known format of these is TBX (Term Base eXchange) which is promoted by LISA, the Localisation Industry Standards Association (*www.lisa.org/standards/tbx*).

1.2 Terminological Entry

TBX models a terminological entry. Such entries are built upon the distinction between concepts (which are semantic units) and terms (which designate such units in different languages). One of the first terminological databases, the TEAM system (Hohnhold 1984), consisted of a meta-language header, covering the concept identification, subject area, term status and other general features, and language-specific sections containing the terms, with denotations, part of speech, definitions, and other language-specific material (see Figure 1).

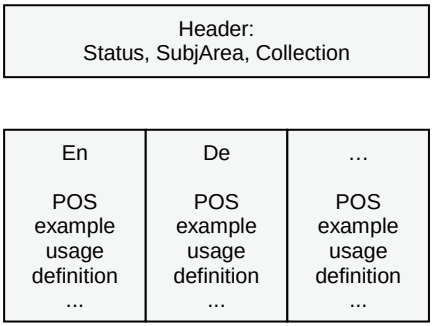

Fig. 1: TEAM database organisation

1.3 TBX Description

The distinction between concepts and terms is still a basic element in the TBX architecture with terminological entries being organised by concepts. Concepts are basic semantic entities; they can have global attributes (stored in the au-

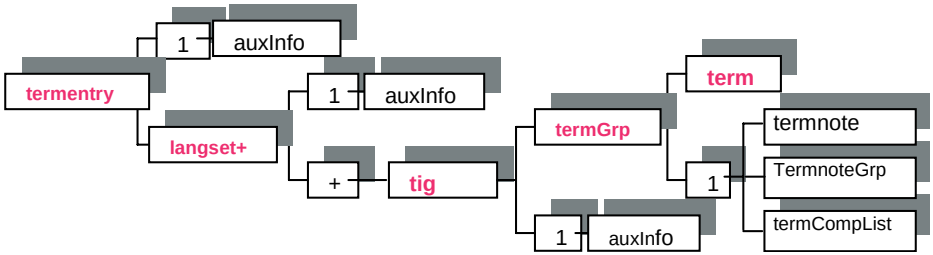

Fig. 2: TBX Term entry structure

xInfo section), like subject area, related concepts, definition, example, sample sentences etc. Then they are described in language-related designation sections (langsets) which consist of term information groups which enclose the single terms. The structure of a term entry is described in Figure 2.

Term entries form the core of a TBX file which is an XML document consisting of the following components (see Figure 3):

A header which describes the file by providing some global and administrative information (content, validation status, contact, encoding, revisions, etc.).

A body which consists of a set of entries, one per concept in the database. The body may have introductory and concluding elements (see Figure 3).

A sample TBX file, taken from the description in *www.lisa.org/standards/tbx*, is given in Figure 4 below.

1.4 Discussion

The meta-model of a TBX entry provides two characteristics:

1. The basic elements of the exchange are concepts, i.e. groups of terms. TBX is based on the distinction between a **concept** (*'Begriff'*) considered to be a unit of thought constituted through abstraction on the basis of properties common to a set of objects; concepts are not bound to particular languages.), and a **term** (*'Benennung'*) considered to be the designation of a defined concept in a special language by a linguistic expression.). As a result, a TBX entry does not consist of single terms but of sets of terms.

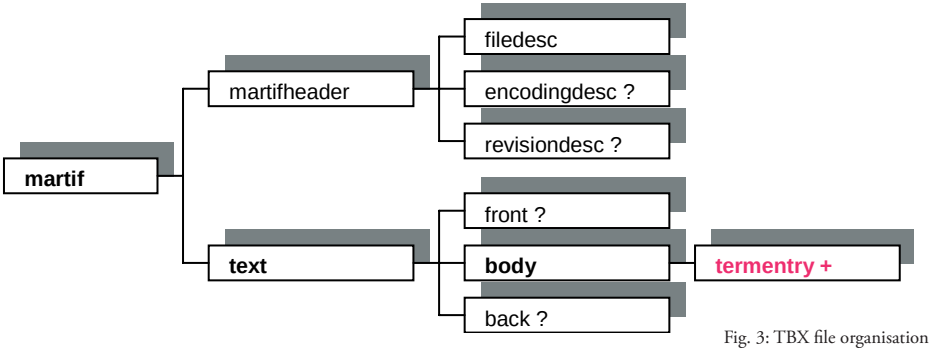

Fig. 3: TBX file organisation

```
<?xml version='1.0'?>
  <!DOCTYPE martif SYSTEM  „./TBXcoreStructureDTD-v-1-0.DTD">
  <martif type='TBX' xml:lang='en' >
  <martifHeader>
    <fileDesc>
      <sourceDesc><p>from an Oracle corporation term-Base</p></
sourceDesc>
    </fileDesc>
    <encodingDesc><p type='DCSName'>TBXdefaultXCS-v-1-0.XML</p></
encodingDesc>
  </martifHeader>
  <text> <body>
    <termEntry id='eid-Oracle-67'>
      <descrip type='subjectField'>manufacturing</descrip>
      <descrip type='definition'>A value between 0 and 1 used in
…</descrip>
      <langSet xml:ang='en'>
        <tig>
          <term tid='tid-Oracle-67-en1'>alpha smoothing factor</
term>
          <termNote type='termType'>fullForm</termNote>
```

Fig. 4: Example of a TBX file (English and Hungarian terms)

2. As a consequence, the meta-model is **multilingual**, i.e. there are as many languages as equivalents are provided, all of which can be interchanged freely; and it is **non-directed**, i.e. from a German-English-French term base all possible bilingual terms can be extracted: German-English, English-German, French-German, etc.

An example is given in Figure 5 which shows two German equivalents for a French entry; they are assumed to be synonyms, both linked to the French term.

2 From TBX to OLIF

The first annotated linguistic dictionaries to be exchanged were Machine Translation resources. However, when starting to work on an exchange format for Machine Translation dictionaries, it quickly became obvious that MARTIF / TBX was not able to satisfy the requirements for exchanging such dictionaries. This is due to the fact that terminology and dictionary entries follow different conceptual lines (Hohnhold / Schneider 1991); but it also follows from inherent problems of the TBX standard.

When exchanging MT information, there were three basic questions to be solved: What are the units of exchange? How can the annotations of the single units be described? How are the relations between them organised?

2.1 Annotations of Exchange Units

The attempt to exchange monolingual MT dictionaries failed rather quickly.

1. The linguistic descriptions available in the TBX standard were not satisfactory for linguistic exchange. In ISO 12620, only very few

```
<termEntry>
  <langSet lang='"fr">
    <ntig>
      <termGrp>
        <term>échantillonneur</term>
      </termGrp>
    </ntig>
  </langSet>

  <langSet lang="de">
    <ntig>
      <termGrp>
        <term>Abtastglied</term>
      </termGrp>
    </ntig>
```

Fig. 5: French term with two German synonyms

annotations are covered (like part of speech, gender, number), and they refer to very few languages. In particular, there was no notion of the basic features to be exchanged in MT, like inflection paradigm, syntactic types, argument structures, semantic features etc., not even complete part of speech sets were provided[1]. As a result, it became clear quite quickly that an extension was required to cover most of the features which an MT system was supposed to exchange.

2. The organisation of the linguistic annotations was not obvious. Some were linked to the concept level (like definitions, examples, relations like broader/narrower term), others, like part of speech (morpho-syntax) or animacy (semantics) were linked to the term level. As a result, semantic information is represented both on concept and on denotation level which is not intuitive. Therefore, it became necessary to define the basic annotations (attributes and their legal values) for the linguistic information to be exchanged. Previous work (e.g. in EAGLES), as well as inspection of existing MT dictionaries, could be used as a reference.

2.2 Relations between units

The attempt to exchange transfer MT dictionaries failed rather quickly as well.

Most MT systems disambiguate 1:n transfers by **tests and actions**, which is shown in Figure 6[2]:

Transfer entries describe language-pair-specific relations between concepts. As TBX is intended to be multilingual and non-directed, there is no possibility to define bi-language directed information as transfer tests. TBX does not provide means to attach information to the links between the denotations in different languages; there are just term information groups relating information to a monolingual term, and information for concepts; options to further qualify

de *ausführen*	(if direct object is of type <person>)	-> en *take out*
de *ausführen*	(if direct object is of type <program>)	-> en *execute*

Fig. 6: Example of a MT transfer entry

[concept: <kill>]		
de *töten*	(standard language)	-> en *kill*
de *umlegen*	(slang)	-> en *bump off*

Fig. 7: Language register

the relations between concepts and denotations, or denotations of different languages do not exist.

This fact does not just relate to MT transfer tests but also to other phenomena, like language register where the same concept has different denotations depending on the register chosen (see Figure 7).

The same holds for other kinds of relations between concepts, like the ones defined in the ISO 2788 for monolingual or multilingual thesauri, or the more elaborate ones as defined in EuroWordNet (Vossen 1999), only limited relations are defined in TBX.

In general, the TBX model assumes that all denotations of a langset are synonyms, and all langsets are equivalents (as shown in the synonym example above, see Figure 5); there is no possibility to qualify such relations in any way.

There is another consequence of a multilingual non-directed approach: In theory it should be possible to **revert** transfers and create an arbitrary bilingual dictionary from such lists. However, this has never worked in practice. Several attempts to create e.g. an English-French dictionary from a German-English-French source failed. As a matter of fact, authors start writing in their native language, and search for equivalents in other languages, which means that such terminology entries are de facto directed, and cannot simply be reverted. Very often, the target equivalents are a bit more general than the source term, (e.g. de *Lichtbogen* -> en *arc*); this fact results in a very specific and improper translation for a rather general term if the entry is reverted.

As a result, in OLIF a data category called `<equival>` was introduced in the transfer section which encodes the degree of equivalence between two words or phrases. Its value indicates whether an entry can be reverted or not. For MT dictionary exchange it is necessary to model the relations between the members of a TBX langset explicitly, moving the basic unit of exchange from a non-lingual set of terms to a monolingual concept/term.

2.3 Units of Exchange

As there is no general mechanism of linking particular source and target terms, there is no means to define equivalents for **general vocabulary** expressions. As such words, like *find, search, restriction* etc. are quite ambiguous, and need to be defined in the context of the respective language, they cannot be stated in a concept – term type manner.

Therefore, TBX cannot be used for exchange of large portions of MT dictionary terms as the majority of MT dictionary entries are general vocabulary terms. TBX does not claim to support general vocabulary terms, and states that the exchange format is intended to support terminology only.

However, the question is which theoretical distinction underlies the fraction of language that is covered by TBX. While it is supposed to cover terminology (as opposed to general language), this does not seem to be the case: Terminology in areas which are subject to societal or cultural influences is not covered either: In the area of the educational system, legal system, social welfare etc., there is no (non-lingual) concept with terms in many languages; very often there is not even a translation available as the underlying phenomenon does not exist in other societies or languages, although the concepts are clearly special-language terms, and match all requirements of being a term (like the Irish *Leaving certificate*, the German *Abitur* or the English *solicitor*). TBX is suitable for the representation of tech-

nical terms where a 1:1 correspondence between participating languages can be assumed.

As a result, TBX is not able to support the exchange requirements of linguistic resources, be it for machine translation, for monolingual applications, for WorldNet type conceptual relations, or any other linguistic tasks. It only covers a part of terminology exchange.

The reason for this fact lies in a **conceptual inadequacy** of the terminological approach: It assumes that there is a concept which has designations in different languages. This idea separates a concept from language, and in turn makes the concept itself a non-language phenomenon, which is not the case: Following Hegel (1807), a concept can only be thought of in the form of a language expression. It is a commonplace since Saussure's *Cours de linguistique générale* that language is a system of signs, and the meaning of a sign is at least partially co-determined by its position in the language system: The consequence is that a non-lingual concept, without being related to other signs of a language system, cannot be defined.

This is the reason why TBX cannot define general purpose words, nor terminology which is defined in language or social specific contexts. This is also the reason why TBX cannot express relations between the terms of two languages, nor assign linguistic descriptions to concepts.

The conclusion is that concepts are monolingual linguistic entities, and must be described in monolingual terms. The consequence is, then, that there must be an explicit relationship between a monolingual entity in one language, and monolingual entities in other languages.

This is the approach which was taken by OLIF: Concepts are monolingual entities, and relations between concepts are modelled explicitly. This approach is bi- or multilingual, and directed. It is more general than the TBX approach, in fact, the relationship of full and reversible equivalents assumed by TBX covers just one specific case of how such a conceptual relationship can be described.

3 The OLIF Format

3.1 History

The Open Lexicon Interchange Format (OLIF) was first defined in an EC project called OTELO. It was intended to enable OTELO partners to exchange sets of MT entries between MT vendors and MT users; one of the objectives was to provide term data (from a term base like SAP-term) for use in MT systems such as Logos or METAL; it included the exchange format itself as well as converters provided by the MT vendors from and into OLIF.

Later versions of the exchange format were developed by the OLIF consortium, members of which included the main MT providers (Systran, Logos, SailLabs, linguatec) and terminology providers and users (Trados, Microsoft, IBM, European Commission, and others). The initative was (and still is) headed by SAP. The current version added a header structure like TBX, provisions for multilingual ontologies, better XML structuring, and several tools and supportive components (McCormick/Lieske 2005).

OLIF is used by major MT users like the European Commission, European Patent Office, SAP, and other multi-vendor MT systems.

As opposed to other standards like EAGLES/MILE (Calzolari et al. 2002), OLIF intended to be pragmatic, and only exchange information which existing MT dictionaries provide, or can make use of. No information which is not (yet) in use, or which is idiosyncratic to a particular system should be included in the standard.

3.2 The OLIF Meta-model

The basic architecture decision of OLIF was to be concept-based (i.e. the basic unit is a semantic entity); but different from TBX, concepts in OLIF are defined for a given language. Concepts form the nodes of an OLIF entry. Between con-

cepts, there are links which point from one concept to another; these links can be monolingual (in case of thesaurus relations) or multilingual (in case of translations). As a result, the metamodel of OLIF can be characterised as follows:

1. It is concept-based but concepts are monolingual and have linguistic annotations.
2. It is multilingual (there can be links from a concept to many target language nodes) but directed (the links have a source and a target, and cannot easily be reverted).

3.3 The OLIF Entries

3.3.1 Key Description

The first challenge is to characterise the entries of exchange. OLIF entries (see Figures 8 and 9) are characterised by four types of information: a canonical form, a language, a part of speech, and a semantic tag[3].

The **canonical form** needs to be described in more detail, to answer questions like Beamter vs. Beamte, automatischer Anlasser vs. automatische Anlasser vs. automatisch Anlasser (multiword terms in particular can be found in many variants).

The **language** is the language in which the entry is defined.

The **part of speech** was defined based on the EAGLES recommendations; in OLIF only open word classes are supposed to be exchanged (as most of the MT systems have their own idiosyncratic view to function words), so noun, verb, adjective and adverb are the categories used. In case

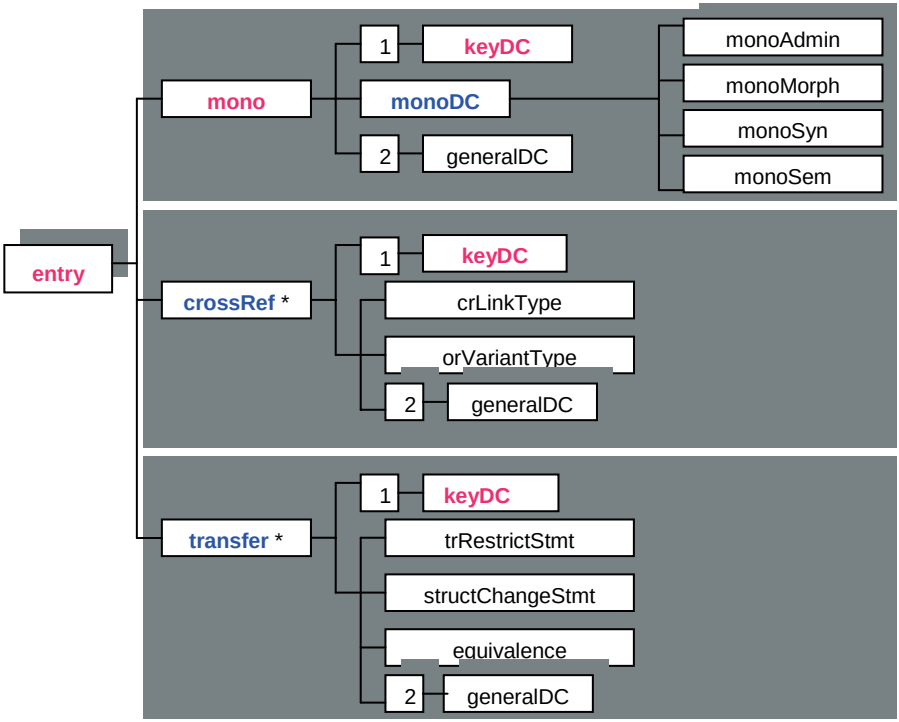

Fig. 8: Structure of an OLIF entry

```
<entry>
  <mono>
    <keyDC>
      <canForm>table</canForm>
      <language>en</language>
      <ptOfSpeech>noun</ptOfSpeech>
      <subjField>general</subjField>
      <semReading>86</semReading>
    </keyDC>
    <monoDC>
      <monoMorph>
        <inflection>like book,books</inflection>
      </monoMorph>
      <monoSyn>
        <synType>cnt</synType>
        <synFrame>[gencomp-opt]</synFrame>
      </monoSyn>
      <monoSem>
        <semType>inform</semType>
      </monoSem>
    </monoDC>
  </mono>
  <crossRefer>
    <keyDC>
      <canForm>row</canForm>
      <language>en</language>
      <ptOfSpeech>noun</ptOfSpeech>
      <subjField>general</subjField>
      <semReading>69</semReading>
    </keyDC>
```

Fig. 9: Example of OLIF entry (from MCCORMICK / LIESKE 2005)

closed classes need to be exchanged, the EAGLES definitions can be used.

The **semantic tag** provides difficulty. In practical situations, the only semantic description available is the **subject field tag**. This can be used as a means for disambiguation. However, there can be cases where the same concept can belong to several subject fields (hand grenade can support both Military and Law-Police subject fields), and several concepts can belong to the same subject fields (like *key* in IT: *Code*, *Schlüssel, Taste*, etc.). Therefore a more precise description had to be found, and a **reading number** was used in addition[4].

3.3.2 Monolingual Annotations

The entry nodes can have linguistic annotations. Such annotations refer to the linguistic and terminological items to be exchanged, and can be grouped according to the levels of linguistic descriptions:

Morphological information is included in most MT lexicons, albeit in very different form as far as inflection information is concerned: While some systems provide different entries for different stems of an inflection paradigm, others enumerate inflection classes for a given stem. As nearly all systems have mechanisms to default the inflection paradigm for unknown words, the idea was to use such components, and use an example-based approach in the exchange standard: Inflection classes are given as simple words, used as examples (`inflects_like`). So, inflection classes in OLIF are Bach, Auto, Haus etc., and it is left to the participating systems to generate their respective internal information structures.

Syntactic information: OLIF provides two basic information types: Syntactic type, i.e. some subcategorisation of the main parts of speech (like the distinction between mass vs. count nouns), and syntactic frames which specify the syntactic argument structure of the respective entries. In particular the argument structure is coded very differently in different MT system, and needs to be converted by each of them from and into OLIF.

Semantic information is also coded. An analysis of the existing MT systems showed that most of them use a simple type system with values such as *human*, *animate*, *place* etc.; more elaborate information is not specified in OLIF as it is not available in such systems.

There is also a section with **administrative information** where author, last editor, validation status and other information is stored.

In general, the objective of OLIF is to cover all such features which make sense to be exchanged. In addition, every system has its own internal information (like: hyphenation information; location in various system dictionaries and the like); this was considered idiosyncratic and did not become part of the standard. Also, information which is relevant but not existing in most systems (like elaborate semantic descriptions) is not part of the standard.

3.3.3 Links

Entries can be connected by links. There are two basic types of links: Links which combine monolingual entries (cross-references), and links combining entries of different languages (translations). Links are directed, i.e. they lead from a source entry (characterised by a key description) to a target entry (characterised by another key description). In addition, links can have attributes:

Crossreferences have a link type (e.g. `is_broader_term_of`, `has_meronym` etc.). The link types have been derived from EuroWordNet; the idea is to support resources used in retrieval, e.g. for query expansion.

Translations have more complex link annotations, consisting mainly of structural descriptions, defining a syntactic configuration (in form of underspecified trees) which must be satisfied for a given link to be activated (i.e. definitions of transfer tests), and
structural changes, defining constellations which define target language changes to be triggered for certain transfers.

It can easily be seen that such attributes of links require them to be directed: Attributes for a German-English link differ significantly from attributes for an English-German entry, even if the same entries are involved. A model like TBX covers only a special case (no link attributes given), and only in this special case a link is reversible.

3.4 OLIF as an Exchange Format

The concept of an exchange format reflects the fact that different systems use different internal representations for dictionary material: For instance, some distinguish between single word and multiword dictionaries, others don't.

As a result, each system participating in the exchange must provide converters from and into OLIF, whereby the proprietary format is converted into the exchange format. Such converters face a number of challenges, given the requirement that conversions should be fully automatic, and complete, i.e. the conversion of a dictionary entry file into OLIF and back should result 1:1 in the same dictionary file:

As OLIF only covers the parts of MT entries which make sense to exchange, there are always **idiosyncratic** parts of the dictionary which are not part of the standard. To be able to exchange the complete dictionary therefore requires system-specific extensions to the standard.

The converters have to cope with all kinds of **mismatches** between the MT systems and the standard definitions:

Proper names are treated as special part of speech in one system and as syntactic subcategorisations in another; mass nouns are considered to be semantic in one system, syntactic in another one. This type of mismatch requires recomputation of the respective values in the converters.

Morphology is a particularly tricky area: Some systems store alternative word stems (for umlauts, irregular verb forms etc.) while the standard only gives numbers of classes, represented by examples. The converters must reconstruct the appropriate information structures when importing OLIF entries.

The **key descriptions** create overhead if dictionary entries are exported and imported for the same system; in this case, the system-internal ID numbers would stay valid, and could easily be used as unique definition marker for a given lexical unit.

The biggest problem in writing converters, however, turned out to be the concept-based **organisation** of OLIF. Most MT systems are lemma-based, and tend to conflate different concepts into one entry to avoid the creation of ambiguities in analysis and parsing. As a result, a syntactic frame like `<Subject – optional Direct_Object>` could describe one concept (with an optional direct object), or two of them (an intransitive one, and a transitive one). If such a concept then has three translations, it is hard to see how the correct assignment of lexical units (with their syntactic description) to translation links could be achieved in a fully automatic way. More research is required to make progress in this area and find a dictionary organisation which keeps the concept-based orientation without giving up the processing advantages of the lemma-based approach.

As a result, writing programs which convert dictionaries from and into OLIF fully automatically and without loss of information is a challenging task.

4 Beyond OLIF

Since OLIF was defined, several other standardisation proposals have been discussed.

4.1. MILE

The MILE standard is the result of research based on EAGLES / PAROLE (Atkins et al. 2002, Ide et al. 2003). It presents the representation of multilingual information in the framework of a layered lexicon representation standard; the morphosyntax being defined by PAROLE, the semantics by SIMPLE, and the multilinguality by ISLE. Unlike OLIF, MILE covers not just the information items which are available in today's MT lexicons but intends to present a complete lexical description, including semantic representation and multilinguality.

In fact, MILE is not an exchange standard but a representation standard, and can be mapped into several different exchange formats as long as they have the expressive power to support all the MILE information categories which holds neither for TBX nor for OLIF.

As it is a layered approach, MILE entries can define overwrite conditions in order to express specific constraints set by the transfer context (e.g. 'target direct object must be in plural') without influencing the monolingual description of an entry.

4.2 XLIFF

This initiative, under the umbrella of the OASIS initiative (*www.oasis-open.org/committees/xliff*), deals with localisation aspects. If a translation job is handed to an agency, usually the text to be translated, the terminology to be used, and the translation memories to be consulted, are delivered in a package. The goal of XLIFF is to standardise the format of such a package: "*The purpose of the OASIS XLIFF TC is to define, through XML vocabularies, an extensible specification for the interchange of localisation information*" (XLIFF V1.1 White Paper 2003).

XLIFF focuses mainly on text handling and translation memory exchange; for terminology exchange, TBX is proposed as standard. There are no specific activities towards terminology exchange.

4.3 Ontology Languages

Ontology languages, the most notable of them being the Web Ontology Language OWL (*www.w3.org/TR/owl-features*), are used to describe meta-information in the context of the semantic web; they describe links between the nodes of the ontology, and rules for derivations and formal properties to be taken care of.

They do not describe any linguistic properties of the ontology concepts, and rarely worry about multilingual issues; the general assumption is that the ontology is a language-independent phenomenon, and each node of the ontology is represented by multilingual terms. This approach is rather similar to the TBX concept.

4.4 Lexical Markup Framework

Recent developments in the effort of standardisation have moved away from the straightforward DTD-based approaches into more general domains of standardisation frameworks, as the efforts for TMF (Terminological Markup framework), supposed to cover both the MATER and the GENETER exchange variants), or for LMF (Lexical Markup Framework) show. The basic idea is to separate two aspects of the exchange formats:

1. The basic data elements to be exchanged, i.e. the **data categories**. This effort, which covers many languages, provides e.g. attributes and values to describe gender, part-of-speech, and other linguistic information items to be exchanged.
2. The way how such data categories can be organised through the provision of **meta-models**; the idea is that implementations like TBX, OLIF and others are just instances of some more abstract meta-model which in turn can cover complete families of exchange formats.

Projects like LIRICS (Linguistic Infrastructure for Interoperable Resources and Systems), an eContent project (*http://lirics.loria.fr/*), and efforts in the context of ISO (TC37/SC4, *http://www.tc37sc4.org/*) try to promote these approaches.

The efforts for standards on data categories could overcome the weakness of the current descriptions in TBX, OLIF, even EAGLES, namely that they support only some of the information categories, and only for some languages. Every time new languages, or new phenomena, are added, the standards need to be revised. A more systematic effort would help to achieve easier exchange of such data, and would also enable resource providers to define easy access using e.g. a common API for morphosyntactic access as defined in LIRICS.

As far as the definition of meta-models is concerned, the challenge is to find a balance between a very abstract model which covers any possible configuration, and proposals which can be implemented and used for exchange of concrete data. It could be proven that both TBX and OLIF can be described with the LMF meta-models, and so could possibly many others.

References

Atkins, S., Bel, N., et al. (2002). "From Resources to Applications. Designing the Multilingual ISLE Lexical Entry". In: Proceedings LREC III, Gran Canaria.

Calzolari, N., Grishman, R., Palmer, M. (2002). "Standards & Best Practice for Multilingual Computational Lexicons: ISLE, MILE and More". In: Proceedings LREC III, Gran Canaria.

Hegel, G. F. W. (1807). "Phänomenologie des Geistes".

Hohnhold, I. (1984). "The TEAM Terminology Data Bank System". In: TermNet News. Journal of the International Network for Terminology, pp. 19-33.

Hohnhold, I., Schneider Th. (1991). "Terminological Records and Lexicon Entries. A Contrastive Analysis". In: META 36, pp. 161-173.

Ide, N., Lenci, A., Calzolari, N. (2003). "RDF Instantiation of the ISLE/MILE Lexical Entries". In: Proceedings ACL Workshop on Linguistic Annotation: Getting the Model Right, Sapporo.

McCormick, S., Lieske C. (2005). "OLIF Tutorial". *http://www.berlinopenforum.de/download/christian_Lieske.zip*

Thurmair G., Lieske, C. (2002). "Lexical Exchange Formats – DXLT and OLIF". In: Proceedings of LRC Workshop, Dublin.

Vossen, P. (ed.) (1999). "EuroWordNet. General Document". EWN Project Report.

XLIFF V1.1 White Paper 2003. *http://www.oasis-open.org/apps/group_public/download.php/3110/XLIFF-core-whitepaper_1.1-cs.pdf*

(Endnotes)

1 Only noun, verb, adjective and other are foreseen.

2 It could be claimed that the example refers to different concepts. This is true and shows that most MT dictionaries are not concept-based. However, even within a concept there can be different translations; this was the starting point to develop language-specific concept hierarchies in EuroWordNet.

3 The idea to characterise an entry by an ID is not sufficient in an exchange format, as both the dictionary where the entries come from and the dictionary where they go to have their own ID systems, and just using IDs in a foreign environment would not really help; an explicit meaning description is required.

4 The definition of an entry on a semantic base raises a huge amount of challenges: How to define it, how to decide on one or several concepts, what about metaphors etc. This is a vast research area, which is explored in lexical semantics, WordNet or FrameNet. However, it is outside of the OLIF standard: Whatever is decided to be a concept can be exchanged in the OLIF format.