Integrated bilingual specialist dictionaries:
The LexTerm initiative

Marie-Jeanne Derouin, Langenscheidt Fachverlag, Munich, Germany
André Le Meur, Université de Rennes 2, France
This paper about integrated specialist bilingual dictionaries in Translation Memories will have two parts:

• **The challenge for the specialist dictionary publishers:**
  - Meeting professional dictionary users' need for global electronic solutions.
  - Integration of their lexicographical contents in language tools such as Translation Memories and automatic Translation machines

• **The methodology for reusing lexicographical data:**
  LexTerm, a bridge between lemma-oriented lexicography and concept-oriented terminology
Actual Translators' needs

- Solutions for optimising their work-flow and time-sparing tools
- A quick access to the language resources they currently use
- A unique tool with « à la carte » integrated specialist dictionaries or other bilingual resources
Future developments of Specialist Dictionary programs

• The current situation: large collections of bilingual specialist dictionaries in technical, scientific and economic fields in print and electronic versions.

• Two main issues
  - Keeping accommodating the needs of a large dictionary users community for traditional printed or electronic dictionaries
  - And meeting the needs of the professional users (Translators and technical writers) for tools beyond Machine Readable Dictionaries on CD-Rom or Online. The specialist dictionary being one component of multifunctional tools for CAT

A joint challenge for specialist dictionary Publishers and TMS providers

- **For Publishers**: An extended data-workflow in order to reuse the existing lexicographical data for the new purposes (A big issue for over a million data)
  - **The challenge**: Two versions of a single source for every specialist bilingual dictionary: a lemma-oriented one for print and electronic dictionaries and a concept-oriented one for integration in other language tools
  - **The Aim**: A better adequacy to the translation market demand

- **For TMS providers**: A new product concept for a unique tool with „à la carte“ specialist dictionaries
  - **The challenge**: the integration of the dictionaries in the translator‘s CAT-tool
  - **The Aim**: An added value for TMS-tools
The solution and the product

• An application of terminological research with data-modelling experts from the Université of Rennes 2

• A longstanding cooperation between Publisher and University:
  – 1998: European project for merging dictionaries and the first DTD for our specialist dictionaries
  – 2000-2006: Major contribution to the new version of ISO 1951. Awareness of the importance of lexicographical standards for reuse and integration of dictionary data

• The product:
  – 10 specialist bilingual dictionaries in over 100 subject fields in the main European languages and in combination with German will be integrated in Translation Memory tools
  – And available on the market in 2007
Application: Integration in Trados
Translator‘s workbench

Ein ausreichender Schutz von Wohnungen gegen Außenlärm ist nur dann gegeben, wenn die Fenster geschlossen bleiben. Die Lüftung muss dann über schalldämmende Lüftungsschlitze in der Aussenwand erfolgen, wobei die Räume an eine gesonderte zentrale Schachtlüftung angeschlossen sind.

Verschiedene Ausführungsarten von schalldämmenden Lüftungsschlitzen (Zuluftschleusen) wurden im Laboratorium untersucht.

Different working plans of sound-insulating

Eine Übersicht über die Schalldämmwirkung der verschiedenen Ausführungen wird gegeben.
across- Integration in Translator’s crossDesk
En-Ge
The editorial workflow

1. Word (RTF document) → Word to XML
   - Lemma oriented XML data
   - XML editor

2. Lemma oriented XML data → LexTerm
   - Concept oriented XML data

3. Concept oriented XML data
   - Terminology Management System
   - Intra/Internet
   - Translation Memory System

- Publisher’s data
- Converters
- Integration by linguistic software providers

LexTerm - Translating and the Computer 28 Conference 2006
LexTerm: a methodology for reusing lexicographical data

- Principle: a bridge between
  - Dictionaries: based on ISO 1951 (XmLex)

- Methodology: semantic and syntactic interoperability
  - Mapping data elements and structures
  - Resolving structural issues
  - Example

- Implementation: an experimental workbench (XML + XSL)
  - LexTermLib an XSL library for transforming dictionaries into monosemic entries
  - A public and online demonstrator
LexTerm: a XSL bridge between standardized XML formats

- XML Standardized dictionary (ISO 1951)
- XML schema
- Semantic repository
- XML schema
- XSL LexTerm converter
- XSL import routine
- XML Standardized terminology (ISO 16642)
- Linguistic tool
Second example: a bilingual general dictionary

- **dam¹ [dæm]** 1 *n a* (wall) *[river]* barrage *m* (de retenue), digue *f* ; *[lake]* barrage (de retenue). 2 *vt a* (also ~ *up*) *river* endiguer ; *lake* construire un barrage sur. **to ~the waters of yhe Nile** faire or construire un barrage pour contenir les eaux du Nil. b **flow of words, oaths** endiguer. 3 *comp dambuster* (*bomb*) bombe *f* à ricochets ; *(person)* (aviateur *m*) briseur *m* de barrages *(se refère à un épisode de la seconde guerre mondiale).*

- **dam² [dæm]** *n (animal) mère f.*
Editing an XmLex entry with XmlSpy

LexTerm - Translating and the Computer 28 Conference 2006
HTML view of an XmLex entry

**dam** [daem]• 1- n. l- (wall) [river] barrage m. (de retenue), digue f.; [lake] barrage m. (de retenue) 2- (water) réservoir m., lac m. de retenue II- ~ up• transitive verb 1- [river] endiguer, [lake] construire un barrage sur | to ~ the waters of the Nile [faire /construire ]un barrage pour contenir les eaux du Nil 2- (flow of words, oaths) endiguer• dambuster l- (bomb) bombe f. à ricochets 2- (person) (aviateur m.) briseur de barrages se réfère à un épisode de la deuxième guerre mondiale

**dam** [daem] n.• (animal) mère f.

**dam** [daem] adj., adv. (vulg.)• 1- damn 4, 4 2- (US) sale [Yankee/Nordiste]
LexTermLib: a XSL library for automatic transformation

15 steps

• Dissociate lemma, multiword units and compositional phrases
• Resolve structural issues
• Deal with synonymy (by grouping referring and main entries)
• Split meanings
• Take morpho-syntactic and pragmatic information apart from semantics
LexTerm
Monosemic entries
The ISO 1951 lemma-oriented meta model

entry

- Headword + morph.syntac. description

Homograph

Sense

- semantic description…

Translation

- Translation description

Compounds

- Compounds description
A lexicographical mark up language : XmLex

• ISO FDIS 1951
  – To be published in 2006-2007
  – A « generic model »
  – Structures + data elements
  – Rules of subsetting
  – Examples
  – Non normative DTDs : XmLex, XmLexForBilingualDictionaries

  ISO 1951 is made for Human Readable and Machine processable dictionaries

It is compatible with LMF (ISO CD 24613) and OLIF (Open Lexicon Interchange Format)
The concept-oriented meta model (ISO 16642 TMF)
Principles of the concept-oriented approach

• Principle and methods: ISO 704
• Data elements: ISO 12620
• ISO 16642: Terminology Markup Framework (TMF)
  - Meta model
  - Terminology Markup Languages (TML), GMT, MSC (=TBX), Geneter (Annex C : Geneter)
The two models

ISO 1951 meta model

XmLex

LexTerm

ISO 16642 meta model

Geneter

entry

Headword + description

Homograph

Sense

Sense description

Translation

Translation description

LIL Concept description

language

LDL Concept description

term

Term description
Methodological aspects of the conversion

- Semantic interoperability (what a data element means)
- Syntactic interoperability (how data elements are combined)
- An example
Semantic interoperability

• Mapping of data elements
  – Common elements (part of speech)
  – Corresponding („mappable“) elements (headword = term)

• Solution: a common semantic repository
  – ISO 11179 model
  – About 2000 elements and permissible values coming from ISO 12620, ISO 1087, ISO 16642, etc.
  
  ... that refers to the ISO TC 37 Data Category Registry
Syntactic interoperability

Mapping structures, taking into account:

- **Synonymy** (referring entries)
- **Homography** (homograph numbers)
- **Polysemy** (sense numbers)
- **Factorization**
Methodology of conversion
An example

A practical case: Langenscheidt specialised dictionaries

1. aerating root s. pneumatophore 1.
2. aerophore s. pneumatophore 1.
3. pneumatocyst 1. (D: Bot) Pneumatozyste f, Luftkammer f (in einem Pneumatophor); 2. s. pneumatophore 2.
4. pneumatophore 1. (D: Bot) Pneumatophor n, Atemwurzel f; 2. (D: Zoo) Pneumatophor n, Schwimmglocke f, Gasflasche f (der Siphonophoren)
First step: Clustering synonyms
Step 2 : Splitting concepts

- **pneumatophore-Botany**
  - en
  - pneumatophore
  - aerating root
  - aerophore

- **pneumatophore-Zoology**
  - en
  - pneumatophore
  - Pneumatophor (n)
  - Atemwurzel (f)
  - pneumatocist

- **pneumatophore-Zoology**
  - de
  - Pneumatophor (n)
  - Schwimmglocke (f)
  - Gasflasche (f)
Syntactic interoperability issues: structural factorization

1 (D: Zoo)
Pneumatophor n, Schwimmglocke f, Gasflasche f (der Siphonophoren)
Syntactic interoperability issues: structural factorization

1 (D: Zoo)
Pneumatophor n, Schwimmglocke f, Gasflasche f (der Siphonophoren)
How to test?

• Read TermBridge home page: http://www.genetrix.org
  (TermBridge is an XML framework for lexicography, terminology, and all their related informations)

• Read XmLex introduction:
  http://www.xmlex.net/lexicography/xmlexintro.pdf

• Download, unpack the LexTermLib.rar:
  http://ww.XmLex.net/lexicography/XmLexWorkbench.rar
Conclusions

- Specialist dictionary publishers will have to act in the future as content providers for language tools in order to meet their actual needs and will therefore have to concentrate more and more on the life cycle of data (production, maintenance, reusability).

- Lexical data bases hosting various semantically and syntactically compatible human-readable formats are the future of linguistic data management based on single sourcing strategies.

- Technically, it is important to rely on publicly available standards (syntax = models, semantics = data elements) and to be compatible with XML and XSL methodology and tools.

- Experience shows that standardization is not a danger for quality and originality of content. In any case it is a guarantee for long life investment.
Thank you for your attention!

a n r e . l e m e u r @ u h b . f r

h t t p : / / w w w . g e n e t r i x . o r g

m a i e - j e a n n e . d e r o u i n @ l a n g e n s c h e i d t . d e

h t t p : / / w w w . l a n g e n s c h e i d t . d e

LexTerm - Translating and the Computer 28 Conference 2006