<u>Minutes</u>

OLIF Meeting for December 5, 2006

The meeting started at 10:30 AM in SAP offices in St. Leon-Rot, Germany.

Participants were:

      Daniel Grasmick, SAP
      André Le Meur, University of Rennes
      Wolfgang Täger, European Patent Office
      Frederik Fouvry, acrolinx
      Gregor Thurmair, Linguatec
      Michael Wetzel, SDL Trados  (by telephone)
      Susan McCormick, Consultant to SAP
      Christian Lieske, SAP
      Véronique Rocca, SAP
      Jennifer Brundage, SAP
      Venkatakrishnan Srinivasan, SAP
      Philippe Bouvard, SAP

- Daniel welcomed participants and gave a brief overview of SAP's interest in/support of OLIF: SAP maintains 1.8 million terms in 31 languages, with up to 120,000 terms/language. Various technologies (MT, term lookup tool, SAPterm), plus sheer size of term base militate for unified exchange option.

- Participants briefly introduced themselves.

- Susan stated the purpose of the meeting as bringing together OLIF users and developers to discuss needs and to review industry support; ultimate goal is to define the direction of future OLIF development.

- Frederik presented an overview of acrolinx's interest in and use of OLIF:

  - acrolinx offers a plug-in language checker, acrocheck, for QA that checks for spelling, grammar, style and terminology (checking and harvesting).
  - Need for a standard exchange format in the work cycle: term harvesting > term import/extraction > term validation > term design > term checking
  - acrolinx OLIF is "central to the terminology component of acrocheck™" and compliant with OLIF 2.1, with acrocheck-specific information stored in a separate namespace; acrolinx supports bidirectional OLIF conversion for term import, export and harvesting.
  - acrolinx has a TBX-OLIF conversion.
  - Suggestions for changes to OLIF that would improve usability for acrolinx include:
    - Support term bank round-tripping: all term bank information must be preserved.
    - Allow representation of deprecated terms (may not have valid alternative term)
    - Support term rules
    - Allow representation of Help info
    - Allow custom-defined data categories
    - Allow representation of options/settings
  - Frederik indicated that several acrolinx customers use TBX and MTF
  - Christian suggested that acrocheck contribute rendering stylesheet OLIF-HTML and bidirectionational converters TBX-OLIF and OLIF-TBX for open use.

- Wolfgang explained how the European Patent Office uses OLIF:

  o EPO selected Worldlingo (Systran-based) MT for automatic translation of patents.
  o Patents organized in classification scheme involving app. 70,000 classifications.
  o MT of "patents/abstracts/communications to/from English,", concentrating on 3 languages/year starting with French, German and Spanish
  o Goal is to improve translation with specific MT dictionaries derived by term extraction in OLIF (done by DFKI).
  o EPO developed OLIF editor for extracted entries to clean up and correct entries and to use a concordancer to provide statistics.
  o Tool needed for distributed dictionary management.
  o EPO requests for changes to OLIF include:
    - Support for more languages (Italian, Swedish, Romanian, Dutch, Greek)
    - Expand/clarify inflection handling and guidelines for canonical forms
    - Explanation of how OLIF maps to concepts found in issue of relational databases vs. flat hierarchy
    - Mechanism to avoid repetition of grammar information for homonym entries

- Christian presented SAP's application of OLIF:

  o SAP uses SAPterm (an SAP application) as Terminology Management System for maintenance; several applications (eg. acrocheck, MT Systems, and terminology lookup tools) are fed with SAPterm data which is exported as OLIF.
  o Srini asked when OLIF will be ready for additional languages. Susan responded that OLIF can already handle this with the concept ID; what will be missing is any detailed grammatical coding for languages that are not included in the original 6.

- André presented his work on LexTerm for specialist bilingual dictionaries :

  o LexTerm is an arbiter between lemma-orientation and concept-orientation.
  o Provides a means of reusing lexicographical data.
  o Has developed 10 specialist bilingual dictionaries for Langenscheidts in 100+ subject areas in the major European languages and in combination with German; dictionaries will be integrated in Translation Memory tools such as Trados' Translators' workbench and on the market in 2007.
  o Basic principle of LexTerm is a bridge between dictionaries (based on ISO 1951 (XmLex, forthcoming) and terminology (based on ISO 16642, TMF (Geneter, Annex C)).
  o ISO 1951 is compatible with LMF and OLIF.
  o An automatic conversion between LMF/Geneter to or from OLIF could contribute to additional reuse opportunities for lexical/terminological data

- Michael (via telecon) updated the group on work on OLIF for MultiTerm:

  o Briefly described differences in usage scenarios for OLIF and MTF (MultiTerm internal format)
  o Noted that MultiTerm format is close to TBX.
  o Have completed development of a plug-in for general conversion framework that converts from bilingual OLIF to MultiTerm, released in late Q1 2007.
  o Noted that OLIF data categories without relevance for MultiTerm users are currently out-of-scope for the converter; guidelines for defaulting (filling mandatory OLIF data categories not present in MultiTerm) may be needed to cover conversion from MTF to OLIF (since MTF may not include OLIF key data categories)
  o SDL Trados is interested in more OLIF test data.
  o MultiTerm will be successor to all SDL terminology applications.
  o Format for date/time in data generated by SAP is inconsistent

- o For SDL Trados KBMT flat file input as starting point for lexicons; extensive post-processing of "seeded" lexicons needed.

- Christian and Susan presented proposals for changes to OLIF for the near future:

  - o Allow markup in certain fields (eg. "definition"), thus enabling the use of XHTML and other formats.
  - o Allow attributes for the data category *subjField;* this will make it easy to use proprietary information re the subject field.
  - o Allow arbitrary strings as values for certain data categories, e.g., subject field; this would enable validation for users where none of the OLIF values for a given data category are relevant.
  - o Change definition of *fileExtent* to correct the existing definition that rules out the use of certain tools (e.g., XMLSpy)
  - o Susan suggested supporting JMdict (a Japanese-specific lexical exchange format already available as an open standard implemented as an XML DTD) in OLIF:
    - JMdict offers a way of addressing the difficulty of mapping headwords in Japanese to OLIF canonical forms.
    - Includes a relatively representative array of grammatical coverage for Japanese entries.
    - Supporting JMdict would be relatively straightforward:
      - Can map overlapping language-general features/values
      - Suggest adding existing JMdict features/values that are general language features to OLIF, e.g., *style*
      - Integrate Japanese-specific features/values perhaps via OLIF extensibility options, i.e., XML namespace

- To close the meeting, there was a general discussion of the viability of the suggested changes and how they could/should be implemented:

  - o Gregor expressed concern about allowing string values for fields like part of speech; Wolfgang and Susan agreed that this pushes the notion of flexibility perhaps to the point of seriously diluting viability as a standard.  A suggestion to offer modules/overlays (using the XSD "redefine" mechanism) to allow for greater flexibity was made.
  - o Daniel and Gregor asked about the ability to cover Japanese requirements for MT. Susan said that the varied requirements of (mostly) proprietary formats for Japanese made JMdict a good option; it is widely recognized and used in Asia for lexical applications and covers many grammatical requirements for MT.
  - o Question was raised about the current status of ISO data category registry 12620 with respect to OLIF.  Susan said that OLIF had provided its data categories for integration into 12620 a number of years ago, but there had been no follow-up since then.
  - o The issue of clarifying/defining extensibility of OLIF for users was raised, e.g., Data Cat Reg vs. XSD vs. namespace.  Christian and Susan will follow up with a formal proposal for the next release of OLIF.
  - o Christian urged participants to seriously consider offering data and tools to the community from the OLIF web site.

The meeting was adjourned at 4:35 PM.

PPT presentations for this meeting are available at:
http://www.olif.net/mediacenter.htm                                        SM, 11.12.06