

OLIF2 Consortium Review Meeting

December 13, 2001

SAP

Walldorf, Germany

MINUTES

The meeting came to order at 9:15 AM.

Present were:

Andrew Bredekamp, DFKI
Deborah Coughlin, Microsoft
Daniel Grasmick, SAP
Michael Kranawetvogl, Bowne
Hubert Lehmann, Linguattec
Christian Lieske, SAP
Susan McCormick, SAP
Peter Quartier, IBM
Gregor Thurmair, Sail Labs
Carlo Mergen, EC
V. Srinivasan, SAP
Michael Wetzel, Trados

1. Daniel welcomed the participants. Susan briefly reviewed the agenda; no changes were made.

New member, Michael Kranawetvogl of Bowne, was introduced. Bowne has MT technology via Globalink and has begun to look at OLIF for exchange purposes.

First-time participant in OLIF meetings, Deborah Coughlin of Microsoft, was introduced. Deborah is an English lexicographer at Microsoft in Seattle.

2. OLIF v. 2 test results were presented by SAP
 - a. Yahoo group site has been established to post test files and messages re test results. Site address is <http://groups.yahoo.com/group/olifConsortium/> for those who haven't yet subscribed.
 - b. Currently, SAP and Systran have posted test results on the Yahoo site:
 - i. SAP has converted SAPterm entries to OLIF v.2; file validated against the DTD with little problem.
 1. The issue of modelling term entries in OLIF arose – concept-orientation, multidirectionality of SAPterm means long, often repetitive entries in OLIF.
 2. It was noted that defaulting strategies already available in the DTD, as well as use of target Ids will help to compress data.
 - ii. SAP has also tested the transfer restriction/structural change formalism; again, the data can be represented, but the entries are large.
 - iii. Systran has posted test entries that also test handling of transfer restrictions. In addition, Systran requires changes to the *morphStruct* element in the DTD in order to model their MT entries.
3. OLIF v.2 applications were presented (ppt files will be posted on the OLIF web site as they are received by the participants)
 - a. Gregor Thurmair discussed issues at Sail Labs resulting from use of OLIF in several tools:
 - i. For term extraction, where output is in OLIF, SAIL would like:
 1. *frequency* as an optional data category

2. elaborated *example* field – currently, OLIF offers just *note*; for bilingual extraction, there are multiple examples in transfer.
 - ii. For OLIF input to the Concept Manager, the modelling of ontological information is difficult. SAIL has moved the *entryStatus* data category to the key level in order to distinguish concept from term entries. Terms are related to concepts by means of cross-references.
 - iii. In general, to accommodate different needs of different applications, SAIL sees ‘different flavours’ of OLIF. This could mean flexibility to create application-specific variants, like those permitted in TBX via the combination of a core formalization (DTD or schema), and an extensible constraint specification file. Christian Lieske pointed out that the data category registry and defaulting strategies in OLIF already afford much of the flexibility that is required.
- b. Hubert Lehmann of Linguattec envisages the possible use of OLIF for generating or importing user dictionaries for Linguattec’s MT. General comments include:
- i. OLIF can be used to generate prose description of entry features for users.
 - ii. The subject field hierarchy is too lean and incorporates no hierarchy.
 - iii. Syntactic frame analysis in OLIF looks like it can accommodate Linguattec; further review of syntactic and semantic values is needed to determine whether OLIF is adequate.
 - iv. Semantic type handling as currently in place is not amenable to Linguattec, which is developing its own lexical semantics.
 - v. Modelling of transfer restrictions will require manual work on part of lexicographers.
 - vi. Suggestion that transfer restrictions be weighted for preference.
 - vii. Mapping of TransLexis format to OLIF should be considered.
- c. Michael Kranawetvogl gave an overview of Bowne technology and discussed how OLIF could be integrated:
- i. Bowne requires coverage of localization features, e.g., *category*, *platform*, *model*
 - ii. Questions on values for inflection – pointed out that all suggested OLIF values are listed on the web; noted that inflections are ‘inflects-like’ type taken from Logos’s inflection pattern tables for all languages but Danish.
 - iii. Suggestion from Bowne that OLIF consortium could develop an editor like the SALT Consortium’s DCS Editor to help map OLIF data categories.
- d. Michael Wetzel of Trados describe the new MultiTerm Client/Server and reiterated that Trados would have OLIF import/export support for MultiTerm. Other comments include:
- i. The OLIF header is heavy; suggestion that we specify fewer header data categories as obligatory.
 - ii. Key data categories may be difficult to generate because data such as part-of-speech may be missing. It was generally agreed that this problem was outside of the scope of OLIF as a format to solve.
 - iii. Trados hasn’t yet decided whether there will be an independent OLIF application or a deeper integration into MultiTerm.
- e. Srinivas of SAP demonstrated some work in progress on an OLIF download of SAPterm data. The download is set for release in February 2002. General observations were:
- i. Approach to OLIF must model the concept-orientation of SAPterm, thus generating large entry constellations in the downloaded file. Data compression and already-existing defaulting strategies were mentioned again as issues/answers for implementation of OLIF.
 - ii. Programming approach has targeted monolingual entries first, to which handling of multilingual or hybrid files can be added.
 - iii. OLIF element names in C constants would be very helpful to programmers.
- f. Andrew Bredenkamp briefly described term extraction work at DFKI in which OLIF is incorporated as an export format.

4. Christian Lieske presented a proposal for approaching validation and certification of applications in which OLIF is used (ppt presentation file to be posted on the OLIF web site). The major points are:
 - a. Certification – certification ensures quality and lends credibility, competitive advantage by declaring that a given item is in conformance with the OLIF standard
 - b. Certification is granted after a conformance assessment that validates, using a test harness, that an item is in conformance.
 - c. Conformance criteria are specifically stated in a conformance clause; conformance criteria indicate different levels of conformance (can consider, for instance, silver/gold/platinum levels of conformance).
 - d. Suggestion to coordinate with SALT’s certification approach.
 - e. Question of copyright/intellectual property and cost of certification process to be addressed.
 - f. Certification can be envisaged for import/export, standalone programs, and files.
 - g. Suggestion to establish working group(s) for:
 - i. Certification Program Policy
 - ii. Conformance Criteria
 - iii. Test Harness
 - iv. Clarification of Business Issues
 - v. Concertation with TBX/LISA/SALT
5. Susan described progress on integrating Asian languages, esp. Chinese, Japanese, and Korean, into OLIF. A working paper by Tom Emerson of new consortium member Basis technology was distributed and discussed. Of special interest was Basis’s recommendation that we replace ISO 639-1 with RFC 3066 for language code use. Tom’s paper is to be posted on the OLIF web site.
6. Susan reported that SALT members had met with Gregor Thurmair of OLIF at the MT Summit in Santiago, Spain in September. Gregor noted that the SALT approach of allowing users to essentially specify their own DTD’s by selecting optional data categories from a data category registry was a model to look at for OLIF. Christian Lieske pointed out OLIF affords comparable flexibility in its current specification, although the structure isn’t modularized in the same way that SALT’s is.

Other points were:

- a. Susan is reviewing the integration of OLIF data categories into the DCR for TBX that was recently distributed by Sue Ellen Wright.
 - b. The most recent TBX specification was distributed for OLIF members to review.
 - c. The general point was made that, while ontologies and concept-oriented models are not easily rendered with OLIF, the lexical model is not easily modelled with TBX; essentially, the OLIF integration into TBX is the availability of OLIF data categories to TBX users.
7. Discussion of future plans led to the following **action items**:
 - ✓ **Target February 1, 2002 for the official release of OLIF v.2. Several minor changes to be made to the DTD/Proposal ASAP by Christian Lieske and Susan McCormick, followed by testing in advance of the release:**
 - a. Add *concept* designation on entry level
 - b. Add *confidence* data category to allow for weighting mechanisms used by term extraction tools
 - c. Look into XLIFF for localization category handling
 - d. Move *fileDesc* to bottom of file to make byte counts easier
 - e. Add data category to allow for specification of default transfer
 - f. Incorporate Systran’s changes to *morphStruct*
 - g. Add an attribute to *note* to specify the type of note
 - h. Consider a *frequency* data category
 - i. Review header data categories to see if some can be made optional

- ✓ Establish working group to begin certification process; Christian Lieske to head.
- ✓ Continue review for Asian languages with an eye to next release of OLIF.
- ✓ Review data compression options; advertise to users already-existing options for defaulting and use of DCR.
- ✓ Consider sending OLIF consultants to help speed the process of integrating OLIF for users.
- ✓ Continue collaborative efforts with SALT; Susan to prepare final report on integration of OLIF data categories into the TBX DCR.
- ✓ Address intractable OLIF DCs (semantic type, subj field, infl, semantic reading) by interfacing with other projects. Gregor will establish contact with SIMPLE and PAROLE for this purpose.
- ✓ Continue testing, especially validated OLIF for MT lexicons
- ✓ Establish plan to advertise OLIF; SAP to consider.

The meeting was adjourned at 5:00 PM

Susan McCormick
December 27, 2001