

OLIF2 Consortium Review Meeting

November 14, 2000

SAP

Walldorf, Germany

MINUTES

1. The meeting came to order at app. 9:20 AM.

Present were:

Gregor Thurmair, Sail Labs
Johannes Ritzke, Sail Labs
Mike Dillinger, Logos
Jean Senellart, Systran
Yves Mahé, Xerox
Carlo Mergen, EC
Michael Wetzel, Trados
Peter Quartier, Lotus
Daniel Grasmick, SAP
Susan McCormick, SAP
Christian Lieske, SAP
Christine Thielen, SAP
Hartmut Bohn, SAP
Laurent Romary, SALT
Markus Pilzecker, DFKI
Andrew Bredenkamp, DFKI
Hubert Lehmann, Linguattec
Anke Holler, IBM
Christel Nahkor, IBM

- Daniel Grasmick opened the meeting and asked participants to introduce themselves and give brief descriptions of their companies'/organizations' interest in OLIF2.
 - The stated purpose of the meeting was 1) "to review, amend and approve proposals for the structure and content of version 1 of the OLIF2 exchange format," and 2) "to plan for the final development, implementation and testing of OLIF2."
2. Christian Lieske presented an overview of the technical group's work on the OLIF2 file header with the focus on two important points: 1) an explanation of the purposes of/rationales behind the header, and 2) options for a formal modelling of the header.

With respect to the purposes/rationales behind the header, Christian noted that:

- Humans/applications need to know if it makes sense to process a certain OLIF2 file at all. For this, several questions must be answered; the header provides these answers.
- The header proposal for OLIF2 has been constructed with an eye to both practical and theoretical considerations. Many data/information categories that have proven useful for other exchange efforts have been included. Mechanisms are in place for allowing an evolution of the format itself, and of tools for processing it. The format to some degree allows a step-by-step approach to implementation.

Regarding formal modeling, he made some comments on suggested changes to the draft version for the header structure:

- The draft version of the formal model for the OLIF2 file header used a standard Document Type Definition (DTD). Since XML schemata provide advantages over DTDs, it was suggested that we use XML schemata for further modelling purposes.
- Input from SAIL Labs on the use of special metadata encoding led to the idea of using the Resource Description Format (RDF) for Metadata.

An ancillary comment concerned the pointers to supplementary, external information that are supported by the current header structure. Christian noted that this facility does not open the door to arbitrary features and values. Rather, it is aimed at allowing the users to add optional, supplementary information of their own as a pragmatic response to gaps that may exist in OLIF2 coverage (e.g. until agreement on OLIF2 ontology has been reached, each OLIF2 file could come with a pointer to an external specific ontology).

⇒ Discussion following Christian's presentation resulted in a number of comments and resolutions for the technical group:

- Both large-scale ideas (the move to XML schemata and the use of the Resource Description Framework) were accepted by the consortium. Laurent Romary from the SALT project indicated his interest in adopting the OLIF metadata ideas in SALT.

**It was agreed that Christian Lieske would integrate any existing SALT ideas; this would be done when the next version of the file header is distributed. SALT would be in contact with Christian to exchange ideas on metadata.
- Several members of the consortium expressed the desire to encode information that is not foreseen as part of the OLIF standard data categories and values. Wrt. to this no final decision has been taken, however
 - pointers to supplementary, external information provide a means
 - the XML namespace mechanism can be used
- Several user representatives in the consortium (e.g. Carlo Mergen from the EC) expressed the need for a data category that supports workflow (this could make use of ISO 12620, 10.3.1).

**The technical group agreed to develop a concrete proposal for workflow support.
- The encoding information will be moved into XML declaration.
- Mike Dillinger and Michael Wetzl became new members of the technical group, joining current members Christian Lieske, Peter Quartier, Gregor Thurmair, and Carlo Mergen.
- The technical group will generate a formal representation of header and body for OLIF2 by the beginning of February, 2001.

3. Gregor Thurmair presented a brief overview of efforts so far to include inflection codes for the six languages covered.

- Original OLIF had inflection class codes for English, German, and Danish
- Attempts to use existing work, e.g., EAGLES, PAROLE have been problematic because they don't represent work that is practically applicable to what we need.

⇒ Recommendation made to use ‘inflects-like’ values for inflection classes.

**Logos has provided its ‘inflects-like’ patterns for English and German, and has said that similar patterns can be provided for French, Spanish, and Portuguese. It was agreed that we would use these for the upcoming version of OLIF2.

OLIF2 should also support user-designated schemas, reference to which is included in the header (e.g., user chooses to use Wahrig’s numbered patterns for German verbs).

4. Susan McCormick briefly reviewed the basic structure of the OLIF2 entry, noting the decision to do away with the allomorph block of features. She then conducted an item-by-item discussion of the data categories that are currently proposed for the body of the OLIF2 file.

⇒ The following agreements were reached re the lexical/linguistic/terminological data categories that were discussed:

- Tag names can be formulated as currently planned – “easy to read, but not too long.”
 - The obligatory *canonical form* should be reinforceable by an optional ‘surface canonical form’ designation; the user can use this to include a transparent, intuitive rendering of the canonical form.
 - The morphological structure of the canonical form should be represented.
 - The value for the *language* tag should simply refer to ISO 639 (i.e., no need for us to define a specific values table).
 - The *language* and *dialect* tags should remain separate, rather than using the combined tags that are now common in terminology handling.
 - The MT developers argued for restricting *part of speech* to open word classes only; some other members were concerned that this would be too narrow. Compromise solution was to specifically support nouns, verbs, adjectives, adverbs, and prepositions, but to also allow for user definition of further parts of speech.
 - It was agreed that the *reading number* value does not, as proposed, represent an objective value that has meaning outside a particular system. Suggestion was made to replace this with a *semantic reading* tag that has general cross-system, cross-linguistic validity.
 - The *entry type* tag should be bifurcated to accommodate the different interpretations of ‘multiword’ between lexicographers and terminologists.
 - Tags such as example, definition, note, usage note should be explicitly described as ‘zero or more’ in the proposal.
 - A values table should be added for the tag *administrative status*; MARTIF can be checked for this; the value *obsolete* should be included.
 - A tag for *orthographic variant* should be added with the value *string*. A type attribute should be accommodated.
- **Linguattec has agreed to provide data on orthographic variant types in German.
- We should get rid of specific administrative tags for the cross-reference and transfer blocks (e.g., crNote) and simply re-use the mono tags for these.
 - The *time restriction* should be explicitly described as ISO-compliant for date and time range.
 - We should consider including tags for derivational morphology.
 - The constituent structure of multiwords should be represented.
- **Systran has offered to propose a formalism for representing multiword structure.
- The *syntactic frame* represents a slot frame analysis. The slots are already in the values table. A syntax for representing the frame is required. Some of the slot value names obviously

imply the filler; question as to whether OLIF2 should also take on the task of defining fillers for the different languages.

- The *preposition* and *particle* tags are used to further specify the syntactic frame. We should make these string fields and let the user specify them freely.
 - Members agreed to leave the *semantic type* values, based on the Logos system, in the specification.
 - A *cross-reference* should be represented as a link type, plus tags for canonical form, part-of-speech, subject area, and reading.
5. Laurent Romary of Loria represented the SALT project for the meeting and agreed to do an impromptu presentation of current SALT work for the OLIF2 consortium meeting.
- XTMF, a format for mapping terminology formats to one another, was described. It includes
 - An underlying representation of term format
 - Structure definition
 - Data registry for data categories from new 12620 data categories
 - SALT will generate XLT automatically using the data categories
 - The TMF site is <http://www.loria.fr/projets/TMF>
 - Suggestion was made to include OLIF2 data categories in the data registry.
6. The issue of inclusion of Japanese into this version of OLIF2 was discussed.
- At present, there is no linguistic support in the consortium to provide the necessary expertise for Japanese
 - It was agreed that this remains a high priority.
- **Consortium members agreed to pursue any possibilities there might be within their own organizations (or with collaborative efforts) to bring Japanese into the project.
7. As a final item on the agenda, plans for the final phase of OLIF2 development were discussed, with the following consortium agreement reached:
- ⇒ The consortium will produce a formal representation of the OLIF2 header and body by the beginning of February, 2001
 - ⇒ A 4-6 week testing phase will begin pursuant to the completion of the formal representation.
 - ⇒ OLIF2 will continue to work in concertation with SALT to ensure as much compatibility as possible.
 - ⇒ The next OLIF2 meeting is tentatively scheduled for April 5, 2001 at SAP headquarters in Walldorf, Germany
8. The meeting was adjourned at approximately 5:00 PM.

Susan McCormick
Consultant to SAP
Baltimore, MD, USA

Christian Lieske
MLT Group, SAP
Walldorf, Germany

December 1, 2000