LIRICS

Deliverable D7.4

Periodic Progress Report M12-M18

Project reference number	e-Content-22236-LIRICS
Project acronym	LIRICS
Project full title	Linguistic Infrastructure for Interoperable Resource and Systems
Project contact point	Laurent Romary, INRIA-Loria 615, rue du jardin botanique BP101. 54602 Villers lès Nancy (France) romary@loria.fr
Project web site	http://lirics.loria.fr
EC project officer	Erwin Valentini
Document title	Periodic Progress Report M12-M18
Deliverable ID	D7.4
Document type	Report
Dissemination level	Public
Contractual date of delivery	M18
Actual date of delivery	30th June 2006
Status & version	Draft
Work package, task & deliverable responsible	WP7, Task7.3, INRIA-Loria
Author(s) & affiliation(s)	Gil Francopoulo
Additional contributor(s)	Adam Funk, Harry Bunt, Thierry Declerck, Monica Monachini, Nuria Bel, Lee Gillam
Keywords	Progress report

Document evolution

version	date	version	date
1.0	30th June 2006		
1.1	14th November 2006		
1.2	15th November 2006		

General progress and deviation

The main actions and events for this period were:

- LIRICS mid-term review
- LREC conference
- Production and CD ballot for LMF

No deviation from the original planning has to be reported.

1 Progress in WP1, infrastructure for standards development and quality assurance

1.1 Progress and deviation

Task 1.1 training and infrastructure for developing standards

Together with partners from DFKI and CNR-ILC, Gil Francopoulo (INRIA) used and tested 'syntax' software that manages the data category registry. This work has been conducted in order to set up a basic set of values for the morpho-syntactic data categories for European languages.

Task 1.3 quality assessment

Mr Neil Newbold took up his post of LIRICS Research Assistant (RA) with effect from 3 January 2006. He has been continuing work undertaken in the first period of the project with specific reference to how *Quality Assurance* relates to document management and to terminologies and terminology products. The work covers the integration and use of supporting resources and components for the standards development process, including: a Plain English thesaurus; ISO TC 37 terminology via ISO 16642; syntactic annotation; and readability metrics. This is specifically relevant to T1.3 and WP5, with the intention of providing an assistive tool to authors of standards based around LIRICS efforts. A round of evaluation has been undertaken between Surrey and INRIA to assess potential for automation of the Plain English task in isolation and the integration of further components and tasks, and the inter-dependency between tasks and standards, will be assessed in the following periods.

1.2 Extended report on activities related to WP1, including LIRICS meetings

A two days meeting (20,21 Feb 2006) has been organized in Paris between DFKI, CNR-ILC and INRIA to work together on the data category registry.

2 Progress in WP2, NLP lexica

2.1 Progress and deviation

Task 1. In the reporting period, CNR-ILC, with the contribution from the partners involved in this Task, has continued the activities concerned with the formulation of a set of Data Categories needed for populating the different layers of the lexical data model being designed in *Task 2*.

In this task, the major outcome is the cooperation between WP2 and WP3 for establishing a harmonized definition of data categories for the lexicon and morpho-syntactic annotation. A meeting was held in Paris (20-21.2.2006) between Lirics/ISO members where a consistent set of morphosyntactic data categories has been decided and inserted in the ISO Morpho-syntactic Profile. As a consequence of this, a new version of the Deliverable 2.1 will be produced where the candidate set of lexical information proposed will be aligned along the different layers of the Morphosyntactic profile (PoS, MorphoFeatures, Case, FormRelated …).

As far revisions coming from ISO experts and suggestions from end-users are concerned, CNR-ILC, thanks to the involvement in international projects, is proposing the emerging set of LIRICS-ISO lexical data categories to NEDO and BootStrep partners in the aim of testing the adequacy and appropriateness and extending their coverage with respect to language-specific considerations (Asian languages) and application (creation of a specialized lexicon for data mining application in bio-medicine domain).

Also related to the activities performed for this task, we can mention the investigation of the possibility of devising harmonized data categories for semantic roles, a topic covered by WP2 and WP4.

Task 2. CNR-ILC has continued the refinement and revision cycles of the lexical Markup Framework, the abstract meta-model for lexical representation. During the reporting period, two versions has been produced: revision-8 and revision-9. On March 2006, the document has been submitted to ISO CD ballot. The ballot lasted three months and on June, 29 the SC-4 P-members voted. Comments coming from voting delegations will be taken into account and discussed during the next ISO Plenary meeting.

As concerns advertising of Lexical Markup Framework to prospective users, a major scientific international conference was held in Genova (Italy), LREC06, where the lexical metamodel has been presented to a larger audience in many different occasions: satellite workshops, main conference and a special dedicated tutorial.

LMF is being adopted in a number of international projects as a model for developing harmonized lexicons in cross-language information retrieval (NEDO project) and for data mining applications (BootStrep project). CNR-ILC, in both projects, is the supervisor for the development of the lexical models ensuring the compliance to LMF. For this purpose, CNR-ILC has implemented an RDF version of the model in order to make partners able to instantiate actual LMF-conformant entries.

Finally, CNR-ILC is designated as the representative of UNI for all standardization activities of ISO/TC37/SC4.

2.2 Extended report on activities related to WP2, including LIRICS meetings

2.2.1 Lirics Meetings:

- Participation of Monica Monachini to the tri-lateral meeting (CNR-ILC, Loria, DFKI) held in Paris (20.02.05 - 21.02.06)

- Participation of Monica Monachini and Claudia Soria to the LIRICS meeting for the preparation of the mid-term review of LIRICS that was held in Genova (in conjunction with LREC06), 23.05.2006. Monica Monachini presented the results of WP2 in the first year of the project.

2.2.2 Events related to LIRICS

- 30.03.2006: Monica Monachini (CNR-ILC) participated to the DIAM Commission in Rome at the premises of UNI, where LIRICS, other standardization activities and involvements in ISO TC37/SC4 has been presented to UNI members.

- Active participation of Nicoletta Calzolari, Monica Monachini and Claudia Soria (CNR-ILC) at the LREC06 Conference (Genova 22-28.05.2006) where a lot of papers about LMF has been presented. Monica Monachini was among the proposers and presenters of the LMF Tutorial which received great success and had a larger audience, with more than 30 participants.

- 23-24.11.2005 in Pisa: organization of a meeting of the sub-group of the ISO/TC 37/SC 4/WG 4 related to MultiWord Expressions and approved by formal resolution adopted at the SC 4 plenary meeting of ISO/TC 37/SC4 in Warsaw, 2005-08-25. The meeting has seen with the participation of Claudia Soria, Valeria Quochi (CNR-ILC) and Gil Francopoulo (LIRICS Technical Manager). Definition of the basic structure of the Extension for NLP related to Multiword expressions.

3 Progress in WP3, morpho-syntactic and syntactic annotations

3.1 Progress and deviation

In this perioding report, DFKI has built on results of Deliverable 3.1 and contributed to both the on-going standardisation activity on Morpho-Syntax (helping in preparing the next CD ballot of MAF) and on elaborating the Working Draft (WD) document for SynAF (SYNtactic Annotation Framework), which has been submitted and accepted as an ISO New Work Item in the last period of LIRICS. A first version of this document has been delivered at the beginning of 2006 and has been discussed at the ISO meeting in Jeju (Korea) between the 19.1 and the 21.1.2006. As a consequence of this meeting, it has been agreed that a new version of the document will be produced, taking into account the comments of the ISO experts present at the ISO meeting in Jeju, and that this version should be distributed to all ISO experts, who have produced a vote on the NWIP of SynAF. This was done in April and comments of the experts have been included in the most recent version of SynAF, which should be then discussed a last time in Paris in July before being presented at the next plenary ISO TC37/SC4 meeting, to take place in August in Beijing. The schedules of SynAF are in line with the plan of the Technical Annex.

DFKI also cooperated with WP2 and WP4 on data categories for the lexicon, the morpho-syntax, the syntax and the semantic. For establishing a consistent definitiion between WP2 (lexicon) and WP3 (morpho-syntax and syntax), a meeting took place in Paris in February 2006, and it was decided to introduce a profile for syntactic annotation The cooperation between WP3 and WP4 (semantics) was also discussed at a more general meeting of ISO TC37/SC4 (TDG3 on the representation of semantic content), which was held with the most relevant American colleagues end of April near Los Angeles. Here we can even say that we are ahead of time considering the plan of the LIRICS TA.

3.2 Extended Report on activities related to WP3, including LIRICS meeting

3.2.1 Lirics Meetings.

- Participation of Thierry Declerck (DFKI) at ISO TC37/SC4 meeting in Jeju, Korea (17-24 January 2006). Discussion of SynAF with Asian and North American representative of standardisation bodies. A lirics meeting takes place on the issue of semantic annotation.

- Participation of Thierry Declerck (DFKI) at the tri-lateral meeting of LIRICS (ILC, Loria and DFKI) 19-21 Februarry 2006 in Paris

- Participation of Thierry Declerck (DFKI) at the joint ISO TC37/SC4, TDG3 and LIRICS meeting in Marina Del Rey: discussing with American partners the issue of semantic annotation, incuding the relation between syntactic and semantic annotation (19-24 April 2006)

- Participation of Thierry Declerck (DFKI) at the LIRICS meeting preparing the mid-term review of LIRICS, that took place on the 23rd of May in Genova (during the LREC conference). Thierry Declerck presents the results of WP at this review.

3.2.2 Events related to LIRICS

- Thierry Declerck (DFKI) worked as a member of the DIN mirroring group NAAT 6

- Thierry Declerck and Mirjam Kessler present LIRICS work at the annual meeting of the DGFS society (21-24.2 2006 in Bielefeld)

- Participation of Thierry Declerck (DFKI) annual meeting of "DIN NAT AA 6 Sprachressourcen", the German Mirror Committee of TC37/SC4 on 12.5.2006.in Berlin. LIRICS was discussed in detail and the DIN committee is very pleased to hear that the works in LIRICS on SynAF and also the semantic annotation is going well. The DIN committeee also work on theCD ballots of LMF and MAF and produces a very exhaustive list of comments, which will be discussed in the ISO TC37/SC4 pleanary meeting in Beiijing (August 2006).

- Thierry Declerck (DFKI) particpates very actively to the LREC conference held in Genova (22nd-28th May 2006).

4 Progress in WP4, semantic content

4.1 Progress and deviation

Following on from the two bilateral meetings between UvT (Tilburg, Netherlands) and DFKI (Saarbrücken, Germany) on 16th November and 14th-16th December 2005 concerning how to isolate a comprehensive set of semantic roles, work in this area began to change from the more theoretical and methodological aspects of semantic annotation towards the development of a more concrete set of data categories for practical annotation. An early attempt at a preliminary list of data categories for temporal and dialogue act annotation was presented and discussed alongside the respective metamodel representations as early as the meeting held at LORIA (Nancy, France) on 10th-12th December 2005. By this point, as discussed in the document D4.1, it had been decided that there would be four areas of consideration for semantic annotation in the data category registry, namely: temporal information, reference, dialogue acts and semantic roles.

In the first half of 2006, data was gathered in earnest on the potential approaches towards a recommendation for a list of semantic roles. The hope was to be able to develop a set of data categories that were as well defined for all areas of consideration. With that view in mind, a literature review was made and some criteria for the consistent definition of semantic roles were formulated. Comparative studies of the various schemes already in existence were carried out. Large areas of overlap were found to exist between different systems of annotation, but little in the way of overall consensus. We wished to identify a rigorous, well-defined and, as far as possible, comprehensive set of data categories that have broad coverage of semantic roles in a variety of different and influential schemes. These data categories would cover what corresponds to the top-level, generic semantic roles, which might then be further refined and specified according to the type of event or action or state defined in the text. These preliminary data categories for semantic roles (along with an updated set of temporal and dialogue act data categories) were presented in document D4.2, and are more or less stable.

The data categories for reference annotation, although also presented in D4.2, are however still under development, and will be subject to considerable further revision pending contributions from other LIRICS participants.

For temporal information a preliminary set of data categories has also been devised. These will serve as input to the newly planned ISO work on the development of an international standard for the annotation of temporal information (see below).

For dialogue acts a set of data categories has been developed that has been discussed at several LIRICS and ISO meetings. In order to design a comprehensible and coherent set of data categories in this area, it was felt important to further develop a multidimensional approach to this aspect of semantic annotation. This approach has been worked out in some detail, has been investigated for its practical use by multiple annotators, and presented at several international conferences. The resulting set of data categories seems stable and extensible.

4.2 Report on activities related to WP4, including LIRICS meetings and dissemination

Other activities during this time frame include:

- ISO/LIRICS meeting in Jeju, South Korea on 19th-21st January 2006 (ISO TC 37/SC 4 meeting where LIRICS work was discussed). Discussion of metamodels for semantic annotation in the areas of LIRICS WP 4; methodological aspects of the design of semantic data categories; and planning of work.

- LIRICS/ISO meeting of experts in semantic annotation in Marina del Rey, Los Angeles, USA on 20th-22nd April 2006:

  1) An intense discussion concerning all areas of semantic annotation (except reference annotation) – mainly planned as a cross-Atlantic workshop to secure greater US support for the development of interoperable semantic data categories.

  2) It was decided to submit a New Work Item Proposal forward to the ISO organization for the development of an international standard for the annotation of temporal information, taking TimeML as starting point, with a project team comprising James Pustejovsky, Bran Boguraev, Harry Bunt, Kiyong Lee and Nancy Ide.

  3) It was agreed to have another joint LIRICS/ISO meeting during the week of the IWCS-7 conference in Tilburg, Netherlands January 2007. Most of those experts present agreed to attend.

  4) There was extensive discussion of semantic roles with Martha Palmer (representing PropBank) and Michael Ellsworth (from FrameNet). There was broad agreement with the idea of the development of top-level data categories (roughly corresponding to the top-level of FrameNet); this has subsequently been implemented.

- LREC conference presentation of Task 4.1. An important opportunity to inform the research community of the effort to produce standards in (particularly) semantic annotation. A lot of interest was shown, especially in the methodological aspects.

5 Progress in WP5, LIRICS reference implementation platform

5.1 Progress and deviation

In the reporting period, WP5 participants have worked on tasks 5.1, 5.2, 5.3, and 5.4.

Task 5.1 worked on further development of APIs (Application Programming Interfaces) following the LIRICS standards for the ISO Data Category Registry (WP1), NLP Lexica (WP2), Morpho-syntactic annotations (WP3). USFD and MPI delivered the LIRICS reference architecture (D5.1.E) in M18.

Task 5.2 has continued and MPI delivered version 2 of the DCR Reference Implementation (D5.2.A). The LMF reference implementation (D5.2.B) has been postponed because a workable XML interchange format has to be developed; MPI and other partners are working on this (which will also involve modifying the existing LMF standard) and it should be ready by M24.

Task 5.3. continued and USFD delivered version 1 of the LIRICS service integration platform (D5.3) at M18. This platform, based on GATE, included a web-service client for morpho-syntactic analysis of Bulgarian and English.

Task 5.4 continued and MPI delivered the first version of the Data Category Usage platform (D5.4) at M18.

Apart from the delay in the LMF implementation, no other deviation from the original planning of WP5 has to be reported.

5.2 Report on activities related to WP5, including LIRICS meetings

DFKI, MPI and USFD met briefly for technical discussions in M17 (at LREC in Genova in conjunction with the project review).

6 Progress in WP6, dissemination and exploitation

6.1 Progress and deviation

According to the fact that a certain number of normative documents are stable now, the partners wrote and presented a great number of scientific communications, in the main conferences of the domain and all over the world.

6.2 Report on activities for WP6 dissemination and exploitation

- 15.02.2066 NEDO meeting in Pisa. Monica Monachini and Claudia Soria presented the LMF meta-model and RDF instantiations of the model has been provided.

- 30.03.2006, DIAM Commission in Rome, at the premises of UNI: Monica Monachini (CNR-ILC) briefly presents ILC-CNR activities connected to standardization, describing the LMF meta-model. The UNI Committee expressed the positive vote on LMF and produced a set of comments that will be object of discussion to the next general ISO TC37 SC 4 meeting in Beijing.

- 13.04.2006 BootStrep kick-off meeting. Monica Monachini presented LMF as the state-of-the-art model for lexicon creation, stressing the importance for a project as BootStrep to work in compliance to this emerging ISO standard. RDF and XML instantiations of the model has been provided to partners.

- 22.05.2005: The LIRICS partners gave a tutorial on LMF during LREC conference. This tutorial was organized by G. Francopoulo, M. Monachini, L. Romary, S. Samon-Alt.

- 22-28.06.2006: LREC, Main conference: Lexical Markup Framework (LMF), G Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria.

- 22-28.06.2006: LREC, International workshop towards a research infrastructure for language resources: The relevance of standards for research infrastructures, G. Francopoulo, T. Declerck, M. Monachini, L. Romary.

- 22-28.06.2006: LREC, International workshop: acquiring and representing multilingual, specialized lexicons: LMF for multilingual, specialized lexicons, G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, C. Soria.

- Thierry Declerck presented the paper by Thirsten Trippel, Ulrich Heid and Thierry Declerck “Standardisierungsarbeit für Sprachressourcen bei DIN (NAT AA 6)}” at DGFS 2006.

- Mirjam Kessler presented a paper by Mirjam Kessler and Thierry Declerck “Data Categories as the link between Language archives and NLP tools” at DGFS 2006

- Thierry Delcerck presented the paper “SynAF: Towards a Standard for Syntactic Annotation” at LREC 2006.

- Thierry Declerck presented a paper by Declerck Thierry, Pérez Asunción Gómez, Vela Ovidiu, Gantner Zeno and Manzano-Macho David, “Multilingual Lexical Semantic Resources for Ontology Translation” at LREC 2006.

- Thierry Declerck presented a paper by Thierry Declerck, Mihaela Vela “Generic NLP Tools for Supporting Shallow Ontology Building” at LREC 2006.

- Thierry Declerck co-authored a paper, in which he stressed the role that can be played by LMF in linguistic descriptions in ontologies: Paul Buitelaar, Thierry Declerck, Anette Frank, Stefania Racioppa, Malte Kiesel, Michael Sintek, Ralf Engel, Massimo Romanelli, Daniel Sonntag, Berenike Loos, Vanessa Micelli, Robert Porzel, Philipp Cimiano *LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies*. In: Proc. of OntoLex06, a Workshop at LREC, Genoa, Italy, May 2006.

- Khurshid Ahmad, Lee Gillam and David Cheng. (2006). Sentiments on a Grid: Analysis of Streaming News and Views. Proc. of 5th Intl. Conf. on Language Resources and Evaluation (LREC), 22-28 May 2006.

- Lee Gillam and Khurshid Ahmad. (2006). Financial data tombs and nurseries: A grid-based text and ontological analysis. Proc. of 1st Intl. Workshop on Grid Technology for Financial Modeling and Simulation (Grid in Finance 2006), 3-4 February 2006.

- Invited talk: Dr Gillam gave the talk "No place for Sentiments?". Presented over *Access Grid* as part of the National Centre for e-Social Science's Access Grid Seminar Series. Audience: University of Surrey; Institute for Advanced Studies - Lancaster University; e-Science North West (ESNW) - Manchester University; Freeman Institute - Universities of Brighton and Sussex (8 June 2006).

- Invited talk: Dr Gillam gave the talk "Sentiment Analysis and Financial Grids" at JISC/ESRC funded workshop on Bridging quantitative and qualitative methods for social sciences using text mining techniques. Hosted by *National Centre for Text Mining* / *National Centre for e-Social Science*, Manchester Conference Centre. (28 April 2006)

- Nuria Bel: LIRICS, Linguistic Infrastructure for Interoperable Resources and Systems”. Poster in the VII Congreso de Lingüística General, Barcelona, April 18-21, 2006

7 Progress in WP7, management

7.1 Progress and deviation

Financial coordination: Virginie Tessier (INRIA) established and maintained the financial records and payments.

Administration and project coordination:

- INRIA with the help of the other LIRICS partners organized the LIRICS mid-term review in Genoa (23rd May 2006).

- INRIA maintained the LIRICS web site and the associated forum where a great number of technical messages were exchanged.

7.2 Extended report on activities related to WP7, including LIRICS meetings

- Gil Francopoulo and Monte George had a two days technical meeting in Washington (10, 11 January 2006) in order to stabilize LMF core model.

- Gil Francopoulo organized the Thematic domain group-2 on data categories for morpho-syntax in Jeju (Korea) and attended the meetings on syntactic annotation and word segmentation (19, 20, 21 January 2006).

- Gil Francopoulo attended WordNet conference in Jeju following ISO meetings.

- In June 2006, Gil Francopoulo has been nominated as French head of delegation for ISO-TC37/SC3 (Terminology and other resources) and ISO-TC37/SC4 (Language resource management).