LIRICS

Deliverable 6.3.B

LIRICS Exploitation Plan

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	LIRICS Exploitation Plan
Deliverable ID	**D6.3.B**
Document type	Report
Dissemination level	Public
Contractual date of delivery	M24
Actual date of delivery	11 July 2007
Status & version	v1.0
Work package, task & deliverable responsible	WP6 (UW), T6.3b (USFD), D6.3.B(USFD)
Author(s) & affiliation(s)	Adam Funk (USFD)
Additional contributor(s)	Gerhard Budin (UW), Gil Francopoulo (INRIA), Thierry Declerck (DFKI), Monica Monachini (CNR-ILC), Marc Kemps-Snijders (MPI), L. Gillam (UniS), Amanda Schiffrin (UVT), Núria Bel (IULA-UPF), Laurent Romary (LORIA)
Keywords	LIRICS Exploitation Plan

Document evolution

version	date	version	date
1.0	11/07/07		

1. Introduction

This document discusses the results foreseen in the LIRICS project and outlines steps for the dissemination and exploitation of those results.

2. Research objectives

To support the spread of localized and multilingual documents today, especially in digital communication, storage and retrieval, LIRICS aims

- to provide ISO-ratified standards for language technology to enable the exchange and reuse of multilingual language resources;
- to provide an open-source implementation platform, related web services and test suites, based on existing formats, tools and data, in order to facilitate the implementation of these standards for end-users;
- to gain full industry support and input to the standards development from the Industry Advisory Group and demonstration workshops; and
- to increase European awareness of language engineering standards and promote their use.

The standards for ratification by the ISO will include:

- a metamodel for lexical representation;
- metamodels and data categories for morpho-syntactic and syntactic annotation;
- reference data categories for semantic annotation;
- test suites in six European languages (Bulgarian, English, French, German, Italian and Spanish); and
- an open-source implementation platform, compatible with certain relevant and well-known systems and tools.

3. Broad dissemination and use intentions for the expected outcomes

3.1. Website

The public section of the LIRICS website (http://lirics.loria.fr/) describes the project's objectives, summarizes the workpackages, and provides all the public deliverables as well as a record (with presentations and other documents) of LIRICS meetings, workshops and other events.

3.2. Standards

LIRICS will build on existing *de facto* standards and initiatives in linguistic annotation (e.g. TEI, EAGLES and ATLAS) to produce standards for ISO ratification (within the lifespan of the project); these will act as common references to facilitate the long-term re-use of language resources and NLP tools. To this end, the project will in particular promote these standards to the relevant industries in Europe.

3.3. APIs and software

The reference implementations and integration platform clients will be published under open-source software licences in order to promote their use and exploitation (for commercial as well as research purposes) beyond the life of the project. The DCR and LEXUS software will be released under an open-source, free-of-charge license for non-commercial use. (The exact details of the license have not yet been determined.)

To ensure availability and encourage community support in the long-term, the source will be made available on well-known servers (e.g. sourceforce) and linked from the GATE and LIRICS websites.

The reference implementations (of services) and clients will conform to the LIRICS web-service APIs, which will themselves be based on XML and language-independent (with regard to natural and programming languages). These APIs will therefore support

- networked distribution of resources and tools;
- provision of large centralized resources (e.g. lexica);
- commercial exploitation by NLP service providers; and
- composition of toolchains by users; and
- future development of further tools compatible with the LIRICS reference suite.

In order to demonstrate how LIRICS standards facilitate the building of multi-component NLP systems and, more importantly, to provide users with means to use LIRICS-compliant distributed resources and build more complex NLP applications based on them, task 5.3 will provide an open-source LIRICS Service Integration Platform. This platform will help users with LIRICS service composition, process flow, and debugging. Rather than starting from scratch, we chose to build on the world-class open-source GATE infrastructure (http://gate.ac.uk). GATE has already a set of graphical tools that support the creation, execution, and debugging of traditional NLP applications, i.e., applications consisting of components executed on the same machine. Therefore, GATE's component model will be extended to support distributed LIRICS services and a new application flow editor will be implemented that enables service composition and debugging. The ACL registry and ongoing relevant work in INTERA will be used to provide the data categories used for describing the NLP tools.

4. General public and scientific dissemination

The APIs and software tools in the reference implementation will be made publicly available from the ISO server and also from the open-source server (http://gate.ac.uk/) maintained by USFD independently of this project. USFD is committed to providing user support and maintaining that server beyond the bounds of the LIRICS project, in the same way as is currently done for the GATE open-source platform.

The open-source reference implementation and the APIs will also make it easier to apply the standards to additional tools and natural languages, as will the use of common implementation platforms and architectures (GATE, Java, Tomcat).

LIRICS acknowledges that many resources still are created manually due to the lack of useful automatic parsers and generators such as for less-spoken languages or for multimodal annotation. In particular in these environments it is important to facilitate the re-usage of existing data categories and the integration of new data categories into the central registry. Task 5.4 will extend the open source world-leading ELAN tool to allow users to easily find and re-use existing Data Categories and to offer a framework to users to formally define new data categories. These could be candidates for enriching the central ISO DCR.

5. Dissemination activities

5.1. Scientific publications

Gillam, Tariq and Ahmad (2005) “Terminology and the Construction of Ontology”. Terminology 11(1), pp55-81. John Benjamins Publishing Company.

Gillam and Ahmad (2005). “Pattern mining across domain-specific text collections”. In Perner and Imiya (Eds.) Proceedings of 4th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM) 2005, Leipzig, Germany, July 2005. Lecture Notes in Artificial Intelligence, Vol. 3587 pp 570-579.

Gillam (2005). “Metadata descriptors: ISO standards for terminology and other language resources”. Proceedings of 1st International e-Social Science Conference. Manchester, June 2005.

Ahmad, Gillam and Cheng (2005). “Textual and Quantitative Analysis: Towards a new, e-mediated Social Science”. Proceedings of 1st International e-Social Science Conference. Manchester, June 2005.

Gillam, Ahmad and Dear (2005). “Grid-enabling Social Scientists: some infrastructure issues”. Proceedings of 1st International e-Social Science Conference. Manchester, June 2005.

Gillam and Ahmad (2005). “Overcoming the Knowledge Acquisition Bottleneck?”. Proceedings of 7th International conference on Terminology and Knowledge Engineering (TKE) 2005.

Ahmad, Gillam and Cheng (2005). “Society Grids”. Proceedings of e-Science All Hands conference.

Ahmad and Gillam (2005). “Automatic Ontology Extraction from Unstructured Texts”. Proceedings of International Conference on Ontologies, Databases and Applications of Semantics (ODBASE) 2005. Lecture Notes in Computer Science (LNCS).

Trippel, DeClerk and Heid (2005). "Standardisierung von Sprachressourcen: Der aktuelle Stand". GLDV Conference, April 2005.

Declerck and Vela (2005) “Linguistic Dependencies for the extraction of domain-specific semantic relations”, Workshop on Biomedical Ontologies and Text Processing, 4th European Conference on Computational Biology (ECCB), July 2005.

Monachini and Calzolari (2005) “Initiatives towards the Integration of Lexicons: MILE is Taking Steps Forwards”, Proceedings of the GLDV Machine Translation Interest Group, Köthen, Germany, June 2005.

P. Wittenburg (2005) Challenges for DRM in Humanities, Challenges Workshop, Amsterdam, November 2005.

P. Wittenburg (2005), Language Archiving at the MPI, International, DELAMAN Workshop, Austin, November 2005.

P. Wittenburg, G. Budin (2005) Standards for Language Resource Management, Language Standards for Global Business Conference, Berlin, December 2005.

Peter Wittenburg, Marc Kemps-Snijders (2005) LEXUS: A flexible web-based Lexicon Tool Interacting with ISO Data Category Registry, KU Study Meeting, Nijmegen, April 2005.

Peter Wittenburg, Marc Kemps-Snijders (2005) Some LIRICS topics, LIRICS Industry Board Meeting, Barcelona.

Peter Wittenburg (2005) Metadata for Language Resources, Open Forum for Metadata Registries, Berlin, April 2005.

Peter Wittenburg, Albert Russel, Peter Berck, Marc Kemps-Snijders (2005) Advanced Web-based Language Archive Exploitation and Enrichment, Language Technology Conference, Posen.

Marc Kemps-Snijders, Peter Berck, Hans Jorgen Bibiko, Albert Russel, Peter Wittenburg, (2006) Language Archive Utilization, DGFS Conference, Bielefeld, February 2006.

Peter Wittenburg (2006) CLARIN – What is this about?, CLARIN Workshop, Paris, Februart 2006.

5.2. Other dissemination activities

Eighth International Open Forum on Metadata Registries, Berlin (Germany), April 11 - 14, 2005

LIRICS Industry Advisory Group meeting, Universitat Pompeu Fabra, Barcelon (Spain), 20-21 June 2005, attended by Systran, ESteam AB, Pearson, Polderland, Sinequa, Thamus, Expert System, Morphologic, HP.

LIRICS IAG Meeting: Thierry Declerck (DFKI) presented the ISO MAF and SynAF initiatives at the LIRICS IAG meeting in Barcelona (21-22 June 2005).

ISO TC37 plenary meeting. Polski Komitet Normalizacyjny (PKN), Warsaw (Poland), 21-26 August 2005. (in the context of NLP, the TC37 plenary meeting is the most important ISO meeting of the year) The LIRICS partners actively participated to the TC37/SC3+SC4 technical meetings.

Language Standards for Global Business. Berlin, Germany, 12-13 December 2005. Presentation of the LIRICS Project by Gerhard Budin and Peter Wittenburg.

Participation of Thierry Declerck (DFKI) at the Lirics Meeting in Paris, March 16-17. Presentation of actual work on extracting morphological relevant descriptors from past and on-going normalization initiatives on morpho-syntactic annotation. The results of the work are already available in a XML Schema. Especially Eagles and Multext East have been considered, but also work done by ISO representatives. Also discussion on the submission of a ISO New Work Item Proposal on syntactic annotation, to be adapted first at the ISO Meeting in Berlin (8-9 April 2005). For this support is provided by INRIA-LORIA.

LIRICS IAG Meeting, Paris, 10 May 2007. Attendance: Bulgarian Academy of Sciences; Business Semantics Ltd.; China National Institute of Standardization; City University Hong Kong; CNR-ILC; DFKI; ESteam; ICT Marketing Ltd.; INFOTERM; INRIA-Loria; Morphologic; MPI; Pearson Education; Sinequa; Systran; USFD; UniS; UTil, UW. Presentations: Overview and history of TC37 activities; 639 family of standards; Morpho-syntactic profile; Lexical Markup Framework; MAF and SynAF; MLIF and SMIL; Controlled Authoring and video annotation; and Reference implementation for DCR, lexica, annotation management.

5.3. Planned dissemination activities

We envisage eContent workshop and training session at a time to be arranged.

The APIs and the reference implementations will be publicly available from the ISO server and the University of Sheffield (http://gate.ac.uk/).

6. Descriptions of each partner's intentions

6.1. INRIA

INRIA is the project coordinator as well as leader of WP1 and WP7. INRIA intends

- to continue working with SFAX university on the building of an LMF lexicon for Arabic;
- to build an LMF syntactic lexicon as a merge of DicoValence, SynLex and LEFFF;
- to update the LMF lexicon Morphalou so it will be compliant with the final LMF specifications; and
- to use SynAF as the format for corpus annotation in the ANR-PASSAGE project, in tools for for manual annotation as well as for producing a 100-Mw SynAF-annotated corpus.

6.2. DKFI

DFKI being a non-profit research organisation, the exploitation of the LIRICS results will not be direclty in commercial terms. But at the same time, DFKI has among others a mission of supporting the fast transfer of technology to the industry, and by now already a lot of Spinoffs have been emerging from DFKI RTD activitites, many of them working directly in the field of language technology. DFKI will promote the results of LIRICS in the fields covered by those companies., and also beyond, in new projects and other industrial partnerships.

On the same level of importance will be the adaptation of the standards to all the tools developed at DFKI within RTD Projects, ensuring also the persistence of resources within the DFKI itself. DFKI is being since an active member in ISO and W3C, and for sure DFKI should be one of the first insitution in adopting results of standardisationwork, which is not only beneficall only for the Language Technology Lab, involved in LIRICS, but for all the labs of DFKI dealing at some level with language resources.

6.3. USFD

The University of Sheffield is interested in scientific dissemination by publishing the reference inmplementations of the services and platforms, as well as papers about them.

Some of the software developed for the SynAF service for English and Bulgarian will be distributed as a plug-in for GATE (the wrapper for the Stanford parser, which can be trained on suitable treebanks to parse additional languages).

6.4. CNR-ILC

With respect to the exploitation of the LIRICS's results, CNR-ILC, which plays the role of reference point for language resource providers/users and in the area of standardization, is devoting efforts to orient activities towards a wider audience.

CNR-ILC has many contacts with research groups in Europe and beyond (including Asia) and industrial organizations, where the LIRICS results can be advertised and transferred.

CNR-ILC is fully aware of the extreme importance of standards and of their huge potential for industry: standards add value to products and services of content providers and SMEs and lower the costs of production, use and customization of language resources.

In this light, CNR-ILC, responsible in LIRICS for the development of lexical standards, during the life-span of the project, as side activity, has tried to experiment methods and tools to make the standards operational and enable existing lexical resources to comply with them. CNR-ILC is trying to push this framework and export it to industrial realities. Industry is the main vehicle for dissemination and success of LIRICS results: the members of the Industry Advisory group are applying/will apply these standards in their products and demonstrate their utility, viability and relevance.

CNR-ILC has obvious interests in making people aware of the importance not only of the meta-models but, above all, of the notion of Data Category and Data Category Registry which will enable to make use of well defined interoperable concepts for annotation and encoding of resources, without re-inventing what already available.

CNR-ILC has also the opportunity to promote the adoption of the LIRICS standards through ELRA, being actively involved in the ELRA Validation Committee, by which the LIRICS standards can be endorsed.

CNR-ILC is porting the ISO-LIRICS lexical standard – LMF – to as much areas as possible, not limited to the NLP community, nor to Europe. CNR-ILC is, indeed, the major promoter of LMF in two on-going projects and is planning to transfer them to other future projects at international level with Asian partners.

The first project is BOOTStrep, where the lexical meta-model is the basis for the development of the model of a large-scale lexico-terminological resource (the BioLexicon) especially designed for text-mining applications in the biomedical domain. This resources has particular features of novelty, the more important one, is that it is the first resource in the sector to be compliant to ISO lexical standards. Thanks to conformity to standards, it accounts for interoperability and extendibility to other areas. Relationships and reciprocal impacts between the two lexical models, the ISO-LIRICS one and the BioLexicon one, go in two directions: the ISO-LIRICS model strongly influences the architecture and the policies of the BioLexicon model, but, vice-versa, the BioLexicon model constitutes both an extension and an implementation of the available standards, thus enhancing the lexical standards itself.

The other reality where the LMF standard plays a central role is the Japanese grant NEDO. This is a project for pushing European lexical standards in Asia and developing harmonized lexical resources for Asian languages. This LMF-compliant NEDO lexicon is expected to support cross-language information retrieval applications, developed by an Asian industrial partner in view of the Olympics, Beijing 08.

6.5. UW

As the main coordinator of dissemination activities, the University of Vienna organizes workshops with industry groups and associations to spread and promote the results of the project.

6.6. UTil

The University of Tilburg is interested principally in scientific dissemination through publishing the results of their semantic annotation work using an annotation scheme based on the core semantic concepts developed during the course of the project.

6.7. MPI

Within the MPI for Psycholinguistics a large language resource archive is maintained and tools are being developed that enable researchers to interact with resources stored in the archive. To promote interoperability between these resources the MPI is dedicated to adopt the results from the standardization efforts both in the archive development strategy as well as in tools development. Our IMDI metadata schema has now part of the Data Category Registry and we have made an active contribution towards incorporating data category selections into our tools. This strategy will be further extended towards other tools as well. LMF is also supported as part of our effort to unify lexica into a common framework. LEXUS is currently the only tool that is capable of handling the fundamental ideas of the LMF framework. As part of our ongoing effort this tool will be expanded and further developed for our user community.

6.8. UniS

The University of Surrey, an academic and research institution, will take the results of LIRICS forward in:

- ongoing evolution of knowledge extraction and management systems developed at Surrey;
- developing new degrees and components;
- new avenues for publication of research; and
- new business opportunities.

Surrey University believes in research-led teaching as a means to provide degrees whose content is current and relevant, and which can be supported by research expertise at various levels. Research interest can be developed by exposure of students to projects such as LIRICS through open talks, through student projects/dissertations, and by the research providing exemplars for further study. The University of Surrey is keen to continue development of degree programme content in computing disciplines. Issues relating to integration and development will be used to improve existing computing modules, in line with the validation expectations of the British Computer Society (BCS) and the Quality Assurance Agency. In the longer term, such improvements could lead to the development of courses dealing specifically with the type of approach to data being promoted by this project. Dealing with metadata, and the LIRICS approach, has already been incorporated into degree programmes at 2 levels: 2nd year undergraduate (Modeling Multimedia Information Retrieval), and Masters level (Challenges for Computing Professionals) with reference to Legal, Ethical and Professional aspects of computing. Scope exists for demonstrating the linkage between LIRICS and e-Science initiatives. Current undergraduate and Masters projects focus on LIRICS-related topics; further topics in the coming years will build on these.

Results of LIRICS will be promoted to industry via the British Standards Institution, and BSI-adopted ISOs (e.g. BS ISO 16642) will be produced subject to full ISO publication. Collaborations with GeoLang Ltd in relation to the data for ISO 639-6 and the World Language Documentation Centre (WLDC) have led to discussions with the OmegaWiki project, a project of the Wikimedia Foundation (Wikipedia, etc), regarding adoption of LIRICS standards for providing an “Open Source” portal to language resources. OmegaWiki aims: “to produce a free, multilingual resource in every language, with lexicological, terminological and thesaurus information”, and adopting standards is a key to this. GeoLang Ltd have specific interests in supporting language resources and are well positioned to make use of the LIRICS standards in service provision. Further research publications and knowledge transfer activities are planned.

6.9. IULA-UPF

IULA-UPF's interest in LIRICS results is primarily in the internal use of standards. IULA is producing resources (corpus, lexica, etc.) which will adhere to the standards as soon as they are accepted by the ISO organization. The availability of interoperable resources enriched with metadata will increase the possibility of their use in complex searches, linking with other resources, and other interesting services. For the same reasons, IULA will propose adherence to LIRICS results in future projects for the development of resources, systems and services, esspecially when transferring technology to the industry.