

Deliverable Identification Sheet

Project ref. no.	22076Y2C2DMAL2
Project acronym	INTERA
Project full title	Integrated European language data Repository Area

Deliverable status	<i>Public</i>
Contractual date of delivery	
Actual date of delivery	
Deliverable number	<i>N/A</i>
Deliverable title	<i>Annual report, all partners</i>
Type	<i>Report</i>
Status & version	<i>Final</i>
Number of pages	<i>10</i>
WP contributing to the deliverable	<i>WP1</i>
WP / Task responsible	<i>ELDA</i>
Other contributors	<i>ILSP, ILC, MPI, LORIA, USAAR</i>
Author(s)	<i>Khalid Choukri, Mahtab Nikkhou (ELDA) , all partners</i>
EC Project Officer	<i>Philippe Gelin</i>

INTERA Annual Report



<http://www.elda.fr/inter/>

The **Integrated European language data Repository Area** (INTERA) is an eContent Project based essentially on two pillars :

- ↙ to build an integrated European language resource area by connecting international, national and regional data centres,
- ↙ to produce new multilingual language resources.

The first goal involves the integration of a critical mass of different types of language resources with the help of metadata descriptions and the interlinking of the resulting distributed resource repository with an existing tool repository, thus enabling users to directly start suitable tools on the included resources. INTERA anticipates that this integrated and interlinked metadata description domain will facilitate the access to language resources in Europe and help professionals in industry, the eContent business, research and education, and increase the usage of the resources already available.

The second goal addresses the lack of quality of multilingual resources, especially for the less widely spoken languages, including Balkan ones, which are of crucial importance to the development of the eContent business. INTERA goes further ahead by developing exemplary methods for their business attractive production.

Summary of 2003 Activities

- Major achievements (e.g. completion of market and user requirements survey, completion of demonstrator design or implementation, initial reactions and feedback from users, results of field test).

The project started in January 2003. The major achievements of the year 2003 work include:

Integrated Resource Domain

- Major progress on Metadata Sets and Tools
- Selection of Data Centres and Resources

Most contracts with Data Centres were signed. In the section in important work area a matrix gives an overview about the resources to be integrated.

Standardised Descriptions

- Drafting of a stable proposal for the representation of metadata information for language resources and tools.
This draft has been submitted to ISO committee TC 37/SC 3 on November 1st as for a three month DIS (Draft of International Standard) for project ISO 12620-1 (Data Category Registry).
- Preparation of the IMDI specification to make it as conformant as possible to the document submitted to ISO, in order to make it the localisation of this specification by the various partners in the project possible.
- Implementation of an on-line tool dedicated to the browsing and selection of metadata descriptors, to allow an international dissemination of the work achieved in the project.

Repository Linking

- Setting up of a first demonstrator implementing the LREP protocol mechanisms:
The prototype shows how distinct types of repositories containing language resources and a repository listing NLP tools can communicate in an efficient way using appropriate metadata.

Resource Production

- Completion of study on user needs on Language Resources.
- Completion of specifications for the production of Multilingual Language Resources.

Evaluation

- Setting up the preparatory and preliminary phase of the evaluation procedure:
The objects to be evaluated strictly influenced the approach and the process of evaluation.
 - Setting up the evaluation methodology
 - Definition of the evaluation tasks
- The final version of the demonstrator implementing the LREP mechanisms will show also the ability to start processes on language data in an automated way, demonstrating thus the potential commercial application of such protocols based on metadata and XML interchanges: searching in a efficient way for appropriate tools that can act on available language data and as a result of the process call providing for an extended and enriched set of data.

Dissemination and Exploitation

The project participants have taken part in a number of international events in order to promote INTERA and the aim of the project.

Among the key events, we can mention:

- **ELSNET/ENABLER Resources Information Infrastructure Workshop**, which took place in Paris, 28-29 August 2003.
- **LangTech 2003**, The European Forum for Speech and Language Technology, Paris, 24-25 November 2003
- **Workshop on Balkan Language Resources and Tools**, Thessaloniki, Greece, 21 November 2003

A list of the events at which INTERA participants have taken part is given in the section “Important work area”

- Finish with a statement on where the project is ‘positioned’ for the next year.

Important work area

- For each substantial and important area of work completed/started, make a short **section**. This should be tailored for the reader rather than being orientated around work package names.

Integrated Resource Domain

Work on Metadata Sets and Tools

- the new IMDI version was demonstrated and explained at various meetings
- the controlled vocabularies were optimised
- special profiles were developed for the Sign Language community and for the Dutch Spoken Corpus
- the IMDI Editor has been adapted to handle special profiles
- the IMDI Browser can handle now special profiles also
- the IMDI search component was adapted to handle special profiles
- the manuals are being extended to contain the full specifications
- a system was designed to allow to handle IPR¹ and to turn IPR requirements into access rights in an efficient manner; this Access Rights management shell operates on a central server in this first phase (a distributed version will be developed under a new contract); the mechanisms are tightly interconnected with the IMDI metadata
- the Access Rights Management system is being developed and will become ready in April; intensive tests will have to be carried out; the IMDI browser was extended to support the Access Rights Management component
- a Tree-Building tool is under construction

¹ It was requested by the reviewers that INTERA should look into IPR matters and offer solutions.

Selection of Data Centres and Resources

The following matrix may give an overview about the resources to be integrated:

Partner	Resource Provider	Resource Names	Resource Type	State
MPI	BAS (D)	Smartcom	Multimodal	being signed
MPI	BAS (D)	Verbmobil etc	Speech	being signed
MPI	Meertens Inst (NL)	various Dutch corpora	texts	being signed
MPI	Florence (It)	Lablita Corpus	speech/text	being signed
MPI	Taalunie (NL)	Dutch Spoken Corpus	speech	done
MPI	MPI	Gesture Corpus	Multimodal	done
MPI	MPI	ESF Second Learner Corpus	speech/text	done
MPI	MPI	PMOLL Corpus	speech/text	done
MPI	MPI	various others	speech/text	done
USAR	DFKI	Negra, Tiger	text	being signed
USAR	U Prag	?	text	negotiations
USAR	U Budapest	?	text	negotiations
LORIA	ATILF	Frantext etc	text	being signed

Standardised Descriptions

The following work have been carried out:

- an ISO TC37/SC4 Metadata requirements document was worked out
- an Editors Team was built and the requirements document was distributed
- a mapping definition between the new OLAC and IMDI versions was started and a schema for different mapping types was worked out
- the formalised IMDI specifications entered in the ISO repository by LORIA are being controlled and corrected
- discussions were started how to describe relations other than hierarchical ones to be contained in the ISO data category repository
- the localisation work of the IMDI set was started

Repository Linking

Fine-grained Description of the basic specifications for the LREP (Language Resources Exchange Protocol) protocol has been proposed. LREP is the communication mechanism that acts between the two types of repositories involved in INTERA: Language Data Repository and Natural Language Tools Repository.

A first version of the demonstrator implementing the LREP mechanisms is available after the first year. It allows the selection a list specific resource in a repository of language data and then to start a search mechanism into the set of metadata descriptions associated with the NLP tools repository. In the GUI a list of possible tools is given, with some

details on the specific descriptors that have been guiding the search in the metadata domain:

Searched for:

- language = English
- language = French
- OS = OS2
- OS = Windows

NeuroServer

authoring tool for virtual representatives

Supported languages: **English**
Supported operating systems: **Windows 95/98, Windows 98, Windows NT**
[Natural Language Software Registry](#) [XML tool description](#)

Fastr

A tool for automatic indexing and for terminology extraction

Supported languages: **English, French**
Supported operating systems: **Linux, MacOS, Solaris, Windows 95/98, Windows NT**
[Natural Language Software Registry](#) [XML tool description](#)

Pulavan

morphological analyzer, parser and a lexical inheritance system for the Tamil language

Supported languages: **English, Tamil**
Supported operating systems: **Windows 3.1x, Windows 95/98**
[Natural Language Software Registry](#) [XML tool description](#)

LexiGraf

a multilingual lexicography software environment

Supported languages: **English, Finnish, French, German, Greek, Italian, Portuguese, Spanish**
Supported operating systems: **Windows 95/98, Windows NT**
[Natural Language Software Registry](#) [XML tool description](#)

Link Grammar Parser

The Link Grammar Parser is a syntactic parser of English, based on link grammar, an original theory of English syntax. Given a sentence, the system assigns to it a syntactic structure, which consists of a set of labeled links connecting pairs of words. As of July 2000, we are releasing version 4.0 of the parser. Among the new features of version 4.0 is a system which derives a "constituent" representation of a sentence (showing noun phrases, verb phrases, etc.).
We have made the entire system available for download on the web. The system is written in generic ANSI C, and runs on all platforms with a C compiler. There is an application program interface (API) to make it easy to incorporate the parser into other applications.
The parser has a dictionary of about 60000 word forms. It has coverage of a wide variety of syntactic constructions, including many rare and idiomatic ones. The parser is robust, it is able to skip over portions of the sentence that it cannot understand, and assign some structure to the rest of the sentence. It is able to handle unknown vocabulary, and make intelligent guesses from context and spelling about the syntactic categories of unknown words. It has knowledge of capitalization, numerical expressions, and a variety of punctuation symbols.

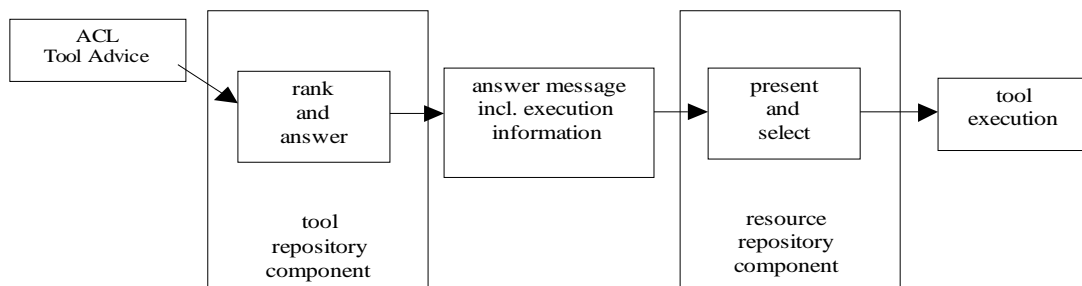
Supported languages: **English**
Supported operating systems: **Linux, MacOS, Solaris, Windows 95/98, Windows NT**
[Natural Language Software Registry](#) [XML tool description](#)

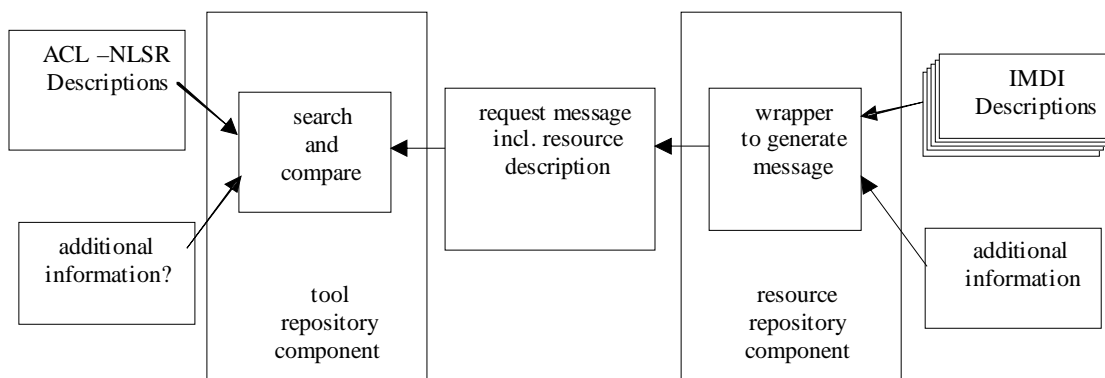
Conexor language analysis tools

<http://www.link.cs.cmu.edu/link/> Trusted sites

The user can then click on the tool he is considering the most appropriate, and the tool can be started in a remote fashion on the data that was the starting point of the search. This remote application scenario in INTERA is limited to two types of tools: morphology and chunking. But this is sufficient for showing the possible commercial application: a user does not longer have to search outside of his/her (metadata) domain for appropriate tools (or related resources): the LREP protocol starts a search in an appropriate metadata domain and proposes solutions. It also ensures the further processing of the data, in a remote fashion or via a downloading to the local storage. The system as such as such can be augment with commercial features like “pay per download” etc.

The following blueprint shows the general mechanism of the LREP protocol behind the demonstrator(a screenshot of which has been given above)





Some of the work described above has been taken over by the subcontractor of USAAR, DFKI, which will be in charge of the transfer of the technology towards prototypes close to marketing. Marketing should be pursued by spin-offs of DFKI.

Resource Production

One of the project's goals is to produce multilingual language resources (MLRs). The specifications for the production of MLRs was based on a study of user needs, aiming to elicit information on the activities of the eContent community related to MLRs (domains and languages of interest, policies concerning acquisition, usage and exploitation of MLRs and tools). This study utilised an especially designed questionnaire, an overview of eContent companies' web sites, and an overview of previous surveys.

Main findings:

- eContent professionals do need MLRs,
- few MLRs freely available over the Internet,
- identification of appropriate MLRs is very difficult,
- development of MLRs in-house or purchase from distribution agencies,
- adherence to standards and availability of related tools are crucial.

The specifications have been finalised and the production of MLRs is ongoing, focussing on Balkan languages in the domains of Health, Law, Education and Tourism.

Evaluation

One of the focuses of INTERA is the definition of an evaluation process aimed at measuring the success of the project. In INTERA, we are concerned with meta-data for language resources and language technology. Being the life cycle of technology in a mature stage of development, the project opted for (i) a user-centred evaluation which relies on expert judgment and user feedback and (ii) an interactive evaluation, where both humans and the system interact. Evaluation will be carried out along two dimensions, (i)

qualitative, where the user satisfaction will be measured on the basis of their expectation and (ii) quantitative, where metrics will be applied to search effectiveness of tools. The evaluation of the content of the produced resources will follow the standard protocols for validation being defined within the ELRA Validation Committee and will be in line with the specifications of the multilingual resources produced during the project and made available to both e-Content and HLT actors.

Central to this approach of evaluation is the definition of the User Group. This group is recruited among e-Content players and digital content producers, key players in HLT development and deployment, users with market in East countries, and translators. The possibility of having two different scenarios of users/evaluators is being allowed for: the first one, where an “Advisory Group” is selected and activated for a kind of pre-evaluation task and a Final User Group is foreseen; the second one, where a unique Final User Group is preferred. Various strategies of recruitment and interaction with them have been envisaged: either personal invitations, letters of recruitment, a flyer to be circulated to the major events of the sector, and advertisement in the INTERA web site, etc. will be used. A Call for Evaluation can be also envisaged.

Consistently to the evaluation criteria defined, a questionnaire has been produced aimed at evaluating the metadata as to their usability, usefulness, completeness etc. Concerning the resource-tool repository interlinking, a discussion is still pending about the possibility of opting for a “free evaluation”, where the user moves freely in the tool or a “task-scenario”, where the user is given a step-by-step task and the questionnaire follows this precise work-flow.

User Group, Promotion and Awareness

The key events in which the participants took part or were involved are the following:

- In order to have an indirect impact on possible market exploitation, INTERA participants participated actively at Standardisation Initiatives, within the ISO framework (TC 37 / SC 4) and within the mirrored DIN German committee (all aspects touching lexical resources, morpho-syntax, and representation of language resources.)
- The INTERA project was presented at the Workshop on Balkan Language Resources and Tools (Thessaloniki, Greece, 21.11.2003), satellite event to the Balkan Conference on Informatics - BCI 2003). The presentation of the project's goals as regards MLRs and Tools raised awareness among the Balkan organisations and resulted in contacts for the production and processing of MLRs on these languages.
- As reported in the former sections, the partners organised **ELSNET/ENABLER Resources Information Infrastructure Workshop**, which took place in Paris, 28-29

August 2003. The connection points between this Workshop and INTERA were, above all, the i) metadata repository and ii) the production of LRs for less spoken languages, crucial topics to be addressed both during the Workshop and the Launching of the International Committee for Written LRs and Evaluation (parallel to COCOSDA). Some of the project partners (Khalid Choukri, Nicoletta Calzolari, Maria Garivaldou, Thierry Declerck) presented some papers at this workshop.

Here after is the list of **conferences and workshops** in which INTERA partners **promoted the project**:

- International E-Meld Workshop, Ypsilanti, July 2003
- International Linguistics Congress, Prague, July 2003
- EUROLAN Summerschool on the Semantic Web and Language Technology, Bucharest, 28 July - 6 August 2003
- DRH Conference, Cheltenham, August 2003
- EUROLAN 2003, July 2003
- Metadata Lecture, Melbourne, October 2003
- Meeting on Bilingual Databases, Nijmegen, October 2003
- Workshop on Balkan Language Resources and Tools, Thessaloniki, Greece, 21 November 2003
- LangTech 2003, The European Forum for Speech and Language Technology, Paris, 24-25 November 2003
- CIT 2003 – Conferencia Internacional de Terminologia, Lisbon 11, 12, 13 December 2003

At all occasions the INTERA intentions were described. At the Prague, ENABLER and LangTech conferences the project flyers were distributed.

Future Work

Regarding the Repository components activities, the demonstrator is making use of up-to date Web technologies that makes it possibly usable in a Webservice scenario. Appropriate Business models are not developed by USAAr, as an academic partner

Future work also covers the actual production of the MLRs, accompanied by a proposal for a business model describing the steps to be followed for the production of MLRs, taking into account the actual problems faced during this effort

In the context of promotion, the partners are organising the **INTERA workshop** (INTERA panel) at **LREC 2004**, Lisbon, 26-28 May 2004. This conference is one of the major event in Language Technologies. LREC 2002 gathered over 700 people from all over the world.

INTERA project participants have submitted several papers for that event which have been accepted. Thierry Declerck is co-organisator of a INTERA-ISO Workshop at this conference.

Further Information

- Inera central website
<http://www.elda.fr/intera>
- Inera hosted by MPI
<http://www.mpi.nl/INTERA/>
- LREC 2004 website
<http://www.lrec-conf.org/lrec2004/>