Intera Deliverable D2.1

Integrated Resource Domain

Project reference number	e-content EDC-22076 INTERA / 27924
Project acronym	*Intera*
Project full title	Integrated European language data Repository Area
Project contact points	Khalid Choukri Evaluations & Language Resources Distribution Agency S.A. 55-57 Rue Brillat Savarin, 75013 Paris, France Phone: +33 1 43 13 33 33 , Fax: +33 1 43 13 33 30 Email: choukri@elda.fr
Project web site	http://www.elda.fr/intera
EC project officer	Philippe Gelin
Document title	Resourse Selection Report
Deliverable ID	D2.1
Document type	Report
Dissemination level	PP
Contractual date of delivery	M3 (01.04.2003)
Actual date of delivery	03.04.2003
Status & version	preliminary
Work package & task ID	WP2, Task 2.1
Work package, task & deliverable responsible	MPI, USSA, LORIA
Number of pages	3
Author(s) & affiliation(s)	Daan Broeder, Kerstin Mauth, Peter Wittenburg, MPI
Additional contibutor(s)	
Keywords	Meta data, Language Resources
Abstract	
Additional notes & remarks	

Resource Selection Report

First efforts have been made in order to identify a set of language resources to be integrated into the prototypical domain. The “subcontracting money” available within the INTERA project will be distributed amongst so-called data providers which will use it to transform or create metadata for their language resources in IMDI format. In work package 2 the MPI and the partners USSA and LORIA are responsible for both the distribution of this subcontracting money as well as for the selection of data providers. Details about the selection procedure are exchanged between the partners mentioned above in order to guarantee a uniform approach to the selection of providers and to the distribution of subcontracting money.

In November 2002 a two-day international workshop was organized at the MPI for Psycholinguistics where representatives of most of the 13 data centers that are the prospective data providers in the project took part (for a complete overview see Appendix 1). At this workshop participants got an introduction to the IMDI metadata set together with a training course on IMDI tools – both developed at the MPI. Furthermore, there were talks by MPI experts in which suggestions for a new version oft the IMDI metadata set for sessions and for lexica were presented. The representatives from the involved data centers each gave short overviews of their language resources and they also tried to sketch the IMDI modifications that might be necessary for integrating their specific resources.
On the basis of the vivid discussions at the workshop, we have already started to create the appropriate documentations and schema for a new version of the IMDI metadata set.

The data centers were contacted again after the workshop in order to get a more detailed overview of the specific language resources especially with respect to size, relevance, complexity, status and format of metadata (if available at all), multilinguality, multmodality etc. They were asked to fill in a resource description form (see Appendix 2) on the basis of which the following overview matrix was created.

Data Centre	Name of Corpus	Corpus Type	Size	MD available	MD format	*Nr. Sessions*	Contact	Res. Open	User Base
LIMSI-CNRS	LIMSI-NICE-1	multimodal speech 2D gestures	62,7 GB	yes	On paper	34	J.-C. Martin	don't know	
BAS	SmartKom	Multi-modal/ Multi-media	100 GB	yes	DB	Several 100	Christoph Draxler	Yes (payement)	industry
ATILF	FRANTEXT	written database	4 GB		easy mapping to IMDI	3350 (texts)	Z. Tucsnak	yes (payment)	Lang. Engineers
	TLFi	lexical database	1,5 GB	yes	header info	100.000 (words)		yes (free)	Linguists
	L'Encyclopédie	text / plates	10 GB		header info	20.8 mio (words)		yes (payment)	General Researchers
	Various dictionaries	computerized dictionaries	2 GB		header info			yes (free)	General Researchers
	Dict. Académie française	dictionary	1 GB		header info			yes (free)	General Researchers
LABLITA	C-ORAL-ROM	speech transcriptions in CHAT	7/8 DVDs	yes	easy mapping to IMDI	500 (sessions) 1.200.000 (words)	M. Moneglia	distribution via ELDA	Linguists
	LABLITA	speech transcriptions in CHAT	10 DVDs	yes	easy mapping to IMDI	257 (sessions) 473.080 (words)		no	Linguists

	Corp. Early Acq. Italian	speech transcriptions in CHAT	?	yes	easy mapping to IMDI	270 (sessions)		no	Linguists
Uni. Helsinki	UHLCS	speech (multimodal?)	20 GB				P. Suihkonen	no	Linguists
Uni. Tübingen	GermaNet	Lexicon		yes	CES header (adapted TEI version)	>4200 sem. concepts	A. Wagner	limited	Linguists
	DEREKO	written database				200 mio. (words)		no	Linguists
	VerbMobil a/b	written database				86.000 (sentences) b) 25.000 (dialogue turns)		no	Linguists
	TUSNELDA	written database						no	Linguists
Meertens Instituut	SAND	written database		no			W. Jongenburger	no	Linguists
	Goeman-Taeldeman DB	speech / transcriptions	25 MB	yes	Filemaker	613 (sessions)		yes	Linguists
	Multimedia Dialect DB	multimodal		no				no	Linguists
Lancaster Uni.	EMILLE	written and spoken		yes	CES header	92.000.000 (various)	T. McEnery	yes	not yet known
CST							Bente Maegaard		
DFKI	NEGRA	written database		no		355.096 (tokens) = 20.602 (sentences)	Thierry Declerck	yes	Computer Linguists

As a second step all prospective data providers were asked to provide samples of the available metadata. This is necessary in order to estimate the work involved to transform the metadata into IMDI format. As soon as the selection process has been finished (but by the end of 2003 at the latest) we will submit a final report about the selection.