META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable 4.3
First upload of language resources

Version No. 1.0
01/12/2011

Document Information

Deliverable number:	D4.3
Deliverable title:	First upload of language resources
Due date of deliverable:	30/11/2011
Actual submission date of deliverable:	01/12/2011
Main Author(s):	Aivars Bērziņš, Imre Bartis
Participants:	All
Internal reviewer:	Tilde
Workpackage:	WP4
Workpackage title:	Cross-national collaboration and Pilot service
Workpackage leader:	UHEL
Dissemination Level:	PU
Version:	1.0
Keywords:	Resources, meta-data
Meta-data model applied to metadata:	the META-SHARE V1 metadata model

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
0.1	24/10/ 2011	Fishbone	UHEL	Imre Bartis	Draft
0.2	30/11/ 11	Review	TILDE	Correction of tables	Draft

EXECUTIVE SUMMARY

This report describes the first upload of language resources planned at M10. The first upload contains metadata descriptions of the resources provided by META-NORD partners and complying with the formats agreed by the META-NET projects. Data provided in this first upload is publicly available at: http://www.meta-nord.eu/index.php?p=first-upload.

Table of Contents

Abbreviations

Abbreviation	Term/definition
LRT	Language resources and tools
DoW	The META-NORD Description of Work document
CC	Creative Commons
TILDE	TILDE SIA (Latvia)
UCPH	Københavns Universitet (Danmark)
UT	Tartu Ülikool (Estonia)
UIB	Universitetet i Bergen Organisasjonsedd (Norway)
UHEL	Helsingin Yliopisto (Finland)
HI	Haskoli Islands (Iceland)
LKI	Lietuviu Kalbos Institutas (Lithuania)
UGOT	Göteborgs Universitet (Sweden)
LRT	Language Resources and Technologies
IPR	Intellectual Property Rights
CLARIN	Common Language Resources and Technology Infrastructure
BLARK	The Basic Language Resource Kit

Table 1. Abbreviations

1. Overall summary of the language resources with metadata

An important aim of META-NORD is to upgrade and harmonize national language resources and tools in order to make them interoperable, within languages and across languages, with respect to their data formats and as far as possible also as regards their content.

A further central aim is the definition of standardized resource and tool metadata and mechanisms for making these metadata harvestable, so that distributed resources and tools can be effectively utilized in language technology applications, both in academic research and in industry.

To describe the metadata the META-SHARE metadata model is used. Based on the META-SHARE metadata model and the technical information regarding data formats that are supported by META-SHARE tool, consortium has created META-SHARE XML for the first meta-data upload. Special attention is devoted to make XML format in full compliance with the technical requirements provided by META-SHARE. This will ensure smooth import of metadata descriptions into META-SHARE when stable version will be released.

In total metadata for 67 resources has been described in XML format. Data provided in this first upload is publicly available at: http://www.meta-nord.eu/index.php?p=first-upload.
This deliverable is related to the deliverable D3.1 “First batch of language resources” which includes descriptions of resources provided for public use by the META-NORD partners.

2. List of language resources metadata

Table below describes the uploaded meta-data by partner. Table indicates the name of the resources and the location of XML schema that is frilly available for public.

Table 1 Summary of metadata

Country	Resource name	Link to the metadata
TILDE	Eurotermbank	https://svn.spraakdata.gu.se/repos/metanord/pub/tilde/LexicalConceptual/tilde_etb.xml
	Lithuanian-Latvian dictionary	https://svn.spraakdata.gu.se/repos/metanord/pub/tilde/LexicalConceptual/tilde_ltlv.xml
	Latvian-Lithuanian dictionary	https://svn.spraakdata.gu.se/repos/metanord/pub/tilde/LexicalConceptual/tilde_lvlt.xml
	Estonian-Latvian dictionary	https://svn.spraakdata.gu.se/repos/metanord/pub/tilde/LexicalConceptual/tilde_etlv.xml
	Latvian-English legislation corpus of Republic of Latvia	https://svn.spraakdata.gu.se/repos/metanord/pub/tilde/Corpus/tilde_leg.xml
	Multilingual dictionary of person names	https://svn.spraakdata.gu.se/repos/metanord/pub/tilde/LexicalConceptual/tilde_personnames.xml

Country	Resource name	Link to the metadata
	Corpus of Latvian literature	https://svn.spraakdata.gu.se/repos/metanord/pub/tilde/Corpus/tilde_lit.xml
UCPH	Danish wordnet, DanNet	https://svn.spraakdata.gu.se/repos/metanord/pub/ucph/LexicalConceptual/DanNet-lex-min.xml
	Copenhagen Dependency Treebank1	https://svn.spraakdata.gu.se/repos/metanord/pub/ucph/LexicalConceptual/CDT1-lex-min.xml
	Copenhagen Dependency Treebank2	https://svn.spraakdata.gu.se/repos/metanord/pub/ucph/LexicalConceptual/CDT2-lex-min.xml
UT	The Estonian Reference Corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/Corpus/METANORD45.xml
	Estonian Treebank	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/Corpus/METANORD50.xml
	Estonian WordNet	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/LexicalConceptual/METANORD-L35.xml
	Corpora of morphologically disambiguated texts	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/Corpus/ESTMORFKORP.gz
	Corpora with shallow syntactic annotation	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/Corpus/METANORD9.xml
	English-Estonian and Estonian-English parallel corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/Corpus/METANORD20.xml
	Semantically disambiguated corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/Corpus/METANORD41.xml
	The database of Estonian verbal multi-word expressions	https://svn.spraakdata.gu.se/repos/metanord/pub/ut/LexicalConceptual/METANORD-L30.xml
UIB	Acoustic database for Danish	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-8.xml
	Acoustic database for Norwegian	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-6.xml
	Acoustic database for Swedish	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-7.xml
	Lexical database for Danish	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-5.xml
	Lexical database for Norwegian	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-3.xml

Country	Resource name	Link to the metadata
	Norsk ordbank, Bokmål	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-9.xml
	Oslo-Bergen tagger	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/ToolsServices/UIB-M10-1.xml
	SCARRIE lexicon	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-11.xml
	Sofietrebanken	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/Corpus/UIB-M10-2.xml
	TRIS Spanish-German parallel corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/Corpus/UIB-M10-12.xml
	Norsk ordbank, Nynorsk	https://svn.spraakdata.gu.se/repos/metanord/pub/uib/LexicalConceptual/UIB-M10-9.xml
UHEL	Finnish TreeBank Grammar Definition Corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/uhel/Corpus/UP10-6.2.xml
	Finnish WordNet	https://svn.spraakdata.gu.se/repos/metanord/pub/uhel/LexicalConceptual/UP10-6.1.xml
	Written corpora of old literary Finnish (Vanha kirjasuomi)	https://svn.spraakdata.gu.se/repos/metanord/pub/uhel/Corpus/UP10-6.4.xml
	Corpus of early modern Finnish (Varhaisnykysuomen korpus)	https://svn.spraakdata.gu.se/repos/metanord/pub/uhel/Corpus/UP10-6.5.xml
	Finnish literature classics (Suomalaisen kirjallisuuden klassikoita)	https://svn.spraakdata.gu.se/repos/metanord/pub/uhel/Corpus/UP10-6.6.xml
	Up-to-date word list of modern Finnish (Ajantasainen nykysuomen sanalista)	https://svn.spraakdata.gu.se/repos/metanord/pub/uhel/LexicalConceptual/UP10-6.7.xml
	Frequency list of words in written Finnish (Kirjoitetun suomen kielen sanojen taajuuslista)	https://svn.spraakdata.gu.se/repos/metanord/pub/uhel/LexicalConceptual/UP10-6.8.xml
HI	Icelandic Parsed Historical Corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/hi/Corpus/UP10_7.1.xml
	Icelandic Frequency Dictionary Corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/hi/Corpus/UP10_7.2.1.xml https://svn.spraakdata.gu.se/repos/metanord/pub/hi/Corpus/UP10_7.2.2.xml
	Parliament Speech Corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/hi/Corpus/UP10_7.3.xml
	Hjal Speech Corpus	https://svn.spraakdata.gu.se/repos/metanord/p

Country	Resource name	Link to the metadata
		ub/hi/Corpus/UP10_7.4.xml
	Pronunciation Dictionary for Icelandic	https://svn.spraakdata.gu.se/repos/metanord/pub/hi/LexicalConceptual/UP10_7.5.xml
	The Saga Corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/hi/Corpus/UP10_7.6.xml
LKI	Modern Lithuanian Dictionary	https://svn.spraakdata.gu.se/repos/metanord/pub/lki/LexicalConceptual/METANORD-L16.xml
	Database of Neologisms	https://svn.spraakdata.gu.se/repos/metanord/pub/lki/LexicalConceptual/METANORD-L19.xml
	Database of Lithuanian Historical Ethnic Place Names	https://svn.spraakdata.gu.se/repos/metanord/pub/lki/LexicalConceptual/METANORD-L18.xml
	Geoinformational Database of Lithuanian Toponyms	https://svn.spraakdata.gu.se/repos/metanord/pub/lki/LexicalConceptual/METANORD-L22.xml
	The Dictionary of Lithuanian	https://svn.spraakdata.gu.se/repos/metanord/pub/lki/LexicalConceptual/METANORD-L23.xml
UGOT	Swedish Wikipedia Corpus	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/Corpus/METANORD-UGOT-C2.xml
	Loan Word Typology list	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.8.xml
	Semantic Information for Multifunctional Plurilingual Lexica	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.7.xml
	Preparatory Action for Linguistic Resources Organization for Language Engineering	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.6.xml
	Swesaurus	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.5.xml
	Swedish FrameNet	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.4.xml
	Examples from the Swedish Associative Thesaurus (SALDO)	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.3.xml
	Swedish Associative Thesaurus' morphology	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.1.xml
	Old Swedish morphology	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.15.xml
	Söderwall's Dictionary of old Swedish Supplement	https://svn.spraakdata.gu.se/repos/metanord/p

Country	Resource name	Link to the metadata
		ub/ugot/LexicalConceptual/UGOT-M10-9.14.xml
	Söderwall's Dictionary of old Swedish	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.13.xml
	Schlyter's Dictionary of old Swedish	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.12.xml
	Dalin's dictionary morphology	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.11.xml
	Dalin's dictionary	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.10.xml
	Swedish Associative Thesaurus	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.2.xml
	Keywords for Language Learning for Young and adults alike	https://svn.spraakdata.gu.se/repos/metanord/pub/ugot/LexicalConceptual/UGOT-M10-9.9.xml

4. Feedback on upload procedure

Taking into account the M10 deadline for the first batch of resources, META-NORD consortium had to work with preliminary version of the META-SHARE. Consortium decided to set up the first node at UGOT and set up additional nodes after the final release of the META-SHARE platform will be released by T4ME in January 2012. The plan was to use META-SHARE node at UGOT for all META-NORD partners to upload their metadata and resources. As a back-up plan, it was decided to develop an alternative repository to harvest XML schemas to make the metadata publicly available from one location.

After receiving META-SHARE version v1.0 in early October (M9) META-NORD partner UGOT tried to set up the META-SHARE node. During the set up process critical bugs were discovered in the META-SHARE platform that made it impossible to use it for the first batch. Discovered bugs were documented and reported to META-SHARE technical help desk.

In particular, the most disconcerting issue was that once a resource file had been edited and saved, it was no longer accessible although it appeared to still be in the system. It was also discovered that the search functionality does not work as expected. One of the major bugs discovered was that the xml import/export functionality was partially successful and the editor was apparently not validating against the xsd schemas but use refactored version of the schemas implemented in the software, not the schema model *per se*.

At the end of November, a new intermediary version of the META-SHARE editor (v1.1) was released. The new release has some improvements over the v1.0 editor and the schemas are provided with the software. However, in the editor v1.1, the schemas provided are named v1.0, but does not match the v1.0 metadata XML used by META-NORD. It was discovered that some changes had been made in metadata schema and the new version had discrepancy with v1.0 scheme. As sufficient description of these changes was not provided, it was decided to continue using the v1.0 schema. Since the META-SHARE editor is not fully functional the META-NORD consortium has set up SVN repository for access to XML metadata for the resources provided in the first batch. This data will be imported into META-SHARE when stable version will released.

The first upload for META-NORD partners is publicly available here:

https://svn.spraakdata.gu.se/repos/metanord/pub

5. Conclusions

The current release of the META-SHARE software provided by the T4ME has serious stability and usability issues and is not yet usable for the first upload of metadata. It is important that the identified issues with usability and reliability are fixed for the next uploads of resources. For META-SHARE to be usable and viable for ambitious tasks of the META-NET network, the platform needs strong improvements to make it reliable and easy to use. META-NORD partners are happy to collaborate in the elaboration of the META-SHARE platform by evaluating it and providing feedback.

The first version of the META-SHARE metadata schemas was stabilized during the first upload. They are available now, together with a draft user manual, and for the first upload META-NORD concentrated its efforts on using the actual metadata model and preparing XML schemas.

To fully move to the META-SHARE metadata model, the metadata editor of the META-SHARE software is vitally important, both for the creation of valid and expressive metadata records from scratch and for the maintenance of the existing metadata records. In upcoming versions of META-SHARE, it is crucial that the editing rights are set up.

Experience in piloting upgrade from the v1.0 to v1.1 showed that current upgrade process is very time consuming, complicated and insuffciciently documented. For the upcoming META-SHARE releases, it is important to ensure that new versions are consistent with the previous ones, clear upgrade instructions are provided and help-desk support provided.