META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure
Project no. 270899

Deliverable D4.1

Metadata descriptions and other interoperability standards

Version 1.0
2011-05-02

Document information

Deliverable number:	D4.1
Deliverable title:	Metadata descriptions and other interoperability standards
Due date of deliverable:	2011-04-30
Actual submission date of deliverable:	2011-05-02
Main Author(s):	Lars Borin, Jonas Lindh
Participants:	Leif-Jöran Olsson, Roberts Rozis, Inguna Sadiņa, Aivars Bērziņš, Martha Dís Brandt
Internal reviewer:	UCPH
Work-package:	WP4
Work-package title:	Cross-national collaboration and pilot service
Work-package leader:	UHEL
Dissemination level:	PU
Version:	1.0
Keywords:	metadata, language resources, language tools, interoperability, standards

History of versions

Version	Date	Status	Author (partner)	Contributions	Description/ Approval level
0.1	2011-02-28	content outline	Lars Borin (UGOT)		
0.2	11-04-01	first draft	Lars Borin, Jonas Lindh, Leif-Jöran Olsson (UGOT)		
0.3	11-04-13	second draft	Jonas Lindh (UGOT)	Martha Dís Brandt (UGOT)	
0.9	11-04-29	final version	Lars Borin (UGOT)	Lene Offersgaard (UCPH), Jussi Piitulainen (UHEL)	Submitted for internal review
0.91	11-05-02	updated final version	Lars Borin (UGOT)	Lene Offersgaard (UCPH), Jussi Piitulainen (UHEL), Martha Dís Brandt (UGOT)	Submitted for final review by Coordinator
1.0	02.05.2011	Final version	Aivars Bērziņš, Inguna Skadiņa, Roberts Rozis		Submitted to PO

EXECUTIVE SUMMARY

Based on a review of the first batch of language resources and tools to be uploaded by META-NORD, this document makes recommendations for a preliminary metadata format and metadata descriptors to be adopted in order to promote interoperability of the resources and tools. It also makes some recommendations on the related issue of resource upgrading, necessary for securing interoperability on the content level.

Table of contents

Abbreviations

Abbreviation	Term/definition
API	Application Programming Interface
CES	Corpus Encoding Standard
CLARIN	Common LAnguage Resources and Technology INfrastructure
CMDI	Component MetaData Initiative <http://www.clarin.eu/cmdi>
CONLL-X format	A CSV format designed for the COmputational Natural Language Learning series of international workshops <http://nextens.uvt.nl/~conll/#dataformat>
CWB	Corpus WorkBench
CSV	Comma-Separated Values
DCMI	Dublin Core Metadata Initiative
DCR	Data Category Registry
DTD	Document Type Definition
HFST	Helsinki Finite State Toolkit <http://hfst.sourceforge.net/>
IMDI	ISLE MetaData Initiative
ISLE	International Standard for Language Engineering
ISO	International Organization for Standardization
ISOcat	ISO TC 37 DCR for widely used linguistic concepts <http://www.isocat.org>
JSON	JavaScript Object Notation
JWS	Java Web Start
LMF	Lexical Markup Framework (ISO 24613)
LOM	Learning Object Metadata
LRT	Language Resources and Tools
MAF	Morphosyntactic Annotation Framework
OAI	Open Archives Initiative
OLAC	Open Language Archives Community
OWL	Web Ontology Language
RDB	Relational Database Management
REST	Representational State Transfer
SOAP	Simple Object Access Protocol
TBX	TermBase eXchange
TC 37 / SC 4	ISO Technical Committee 37 (*Terminology and other language and content resources*) / Sub-Committee 4 (*Language resource management*)
TEI	Text Encoding Initiative <http://www.tei-c.org>
WSDL	Web Services Description Language
XCES	XML Corpus Encoding Standard
XML	eXtensible Markup Language

Table 1 Abbreviations

1. Background

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.

Since resources and to some extent tools will remain in one location – one of a number of META-NORD centers – the preferred way of accessing and utilizing resources and tools will be through *metadata* and *APIs*, allowing the assembly of on-the-fly toolchains made up of standardized component language technology tools, processing distributed – and in many cases interlinked – language resources in standardized formats.

As a consequence of this a further central aim of META-NORD is the definition of standardized resource and tool metadata, standardized tool APIs, and standardized mechanisms for publishing and making the metadata harvestable, so that distributed resources and tools can be effectively utilized in language technology applications, both in academic research and in industry.

The purpose of this document is to make a set of recommendations for the resource and tool metadata to be used in META-NORD. In the META-NORD work package 4 (*Cross-national collaboration and pilot service*) the following goals are set:

> The [META-NORD] consortium will agree on standardized top-level resource descriptions (metadata) for all relevant types of resources, based on a recommended set of metadata for documenting resources provided by META-NET [...]. It will produce such descriptions for each and every resource contributed to the shared pool. Metadata sets will include mandatory as well as optional elements, together with sets of recommended values whenever possible and appropriate. Metadata will include at least information for the resource per se, its identification (including a persistent identifier), together with its creation, annotation, provenance, documentation, usage, availability, licensing and distribution data. There will also be provenance information for the metadata items themselves [...].
>
> Resources will be documented by means of metadata descriptors suggested by and agreed with META-SHARE. A minimum set of Dublin Core (DC) compliant metadata will be compulsory, while extended sets will be used if and when needed. In case existing resources are described using proprietary but popular sets, the consortium will upgrade them using converters, mappers and other tools provided by META-SHARE, or in some cases developed [by META-NORD].

The present document is deliberately being published very early in the project (at the end of project month 3) with the aim of defining a preliminary set of metadata guidelines designed to cover in the first instance the first batch of META-NORD resources to be made available at the end of project month 10 (see section 2 below). So far no recommendations have been published specifically pertaining to META-NET/ META-SHARE. Together, these two circumstances have some immediate consequences for this document and the recommendations given here:

(1a) The present document may need to be published in an updated version when the META-NET/ META-SHARE guidelines become available; and/or

(1b) the present document may need to be provided in an updated version in order to cover the second and third batches of META-NORD language resources to be published later in the

project (at the end of project month 18 and 24, respectively), or as a result of concrete issues encountered while upgrading the first batch resources.

(2) For the time being, we will rely on the work conducted in CLARIN to provide us with best practices and guidelines with respect to formats for language resources, language tools and metadata.

The second point is motivated by the fact that to date no more thorough general recommendations for language resource metadata exist than those defined by CLARIN (see section 1.1), and it is very likely both (i) that META-NET/ META-SHARE will adopt either the CLARIN metadata or some compatible metadata set, and (ii) that the CLARIN metadata will be sufficient to cover the requirements arising from the second and third batch of META-NORD resources.

Good metadata are a necessary but not sufficient requirement for resource and tool interoperability. The data format and – most importantly – the content model of resources and tools must also be standardized for interoperability to be possible. This issue is briefly touched upon in section 1.3 below and must by necessity play a role in the recommendations that are the conclusion of this document: Since the efficacy of the metadata is strongly dependent on the format of the resources, it will be necessary to say something about the recommended upgrading path for the META-NORD resources in order for them to be effectively shared within the META-SHARE framework. This is not surprising, since arguably the content model *is* a kind of metadata, only at a finer level of granularity than the whole resource, and should consequently be included in any discussion of resource metadata and recommendations ensuing from this.

1.1. Metadata standards

For an overview and discussion of relevant metadata initiatives, we refer to CLARIN deliverable D2.4, where a large number of relevant initiatives are listed: METS, OAI-PMH/ OAI-ORE, Dublin Core, TEI, IMDI, Universal catalogue, OLAC, MPEG7, ISOcat, DCR, Natural Language Software Registry, ACL Data and Code Repository, LOM. (D2.4, section 4.2.).

The CLARIN Metadata Initiative can be seen as building on top of the relevant initiatives previously mentioned. The initiative has since been renamed to the *Component Metadata Initiative* (CMDI) since it now aims to become an ISO standard. The data categories, e.g. ISOcat, are the main concern of standardization, not the metadata schema per se. (CLARIN deliverable D2R-5b). The sharing is done by publishing profiles which use components (sets of metadata elements) defined in Data Category Registries (DCRs). CMDI subsumes DC and OLAC.

For more information about CMDI see <http://www.clarin.eu/cmdi> and the FAQ <http://www.clarin.eu/faq/244>.

Examples of already available CMDI metadata converted from earlier formats: IMDI, OLAC. Also available are metadata for the CLARIN LRT Inventory and some early harvested data from CLARIN centres. All these examples can be found here: <http://www.clarin.eu/page/3312>.

The Arbil metadata editor is a tool that can be used for producing CMDI metadata: <http://www.lat-mpi.eu/tools/arbil/>.

1.2. Interoperability standards for language resources

The representation of language resource content has at least two aspects:

(1) There will always be a *data format* for a resource. This is a complex notion involving such components as a character representation (nowadays generally and unproblematically Unicode in utf-8), a model for the structure of the data (e.g., the tabular structure of an SQL database or the hierarchical structure of an XML document). It is important to keep in mind that giving just the data format of a resource is not very informative. Saying that a resource is encoded in XML is a bit like saying that a text is written in English. Of course this provides valuable information about the text, but it also leaves many important things unsaid: What is the subject? Which genre? How difficult is the text? etc.

(2) In the same way, with language resources we would also like to know the *content model*, i.e., roughly the semantics or interpretation of the data: How does a particular SQL table column or a particular XML element or attribute correspond to entities in the domain, etc.?

For interoperability, the content model is the crucial part of language resource representation, since data models can largely be made mechanically interconvertible, whereas, for all practical purposes, content model mapping requires a fair amount of manual effort.

In practice, those working on language resources have converged on a few data formats, of which XML, CSV and JSON are the most important.

Much less progress has been made in the area of content models. There is much ongoing work under the aegis of ISO on defining content models for common types of language resources:

– Lexical Markup Framework (LMF)

– Linguistic Annotation Framework (LAF)

– Morphosyntactic Annotation Framework (MAF)

– Feature structures (FS) 1 and 2

There are also some de facto content model oriented representation standards developed by various organizations:

– TEI

– XCES

– TBX

– OWL

– CONLL

– CWB-type formats

We will see below that several of these data formats and content representation standards are used by META-NORD partners for their resources (Section 2).

1.3. Interoperability standards for language tools

If much has been done already in the area of metadata, data and content representation formats for language resources, this is much less so for language tools, especially with respect to metadata.

For a distributed language resources and tools infrastructure such as the one envisioned in META-NORD and META-NET to work, with ad-hoc toolchains formed from existing tools residing in distributed repositories, the existence of interoperable tool APIs and metadata will be crucial.

At present, initiatives such as CLARIN seem to be focusing on web services as tool APIs (CLARIN deliverable D2.R6). In this framework, a tool is invoked via a web connection using either standard http commands (REST) or a special XML format (SOAP). In both cases, tool metadata can be specified using an XML-based metadata format (WSDL). Data is transferred among web services in a variety of data formats (in the sense used above), with JSON rapidly gaining popularity.

2. Existing resources, tools and metadata in META-NORD: requirements for standardization

As already mentioned in the introduction (section 1 above), one of the main rationales for publishing the present report early in the project was the need to define preliminary metadata for the first batch of META-NORD language resources and tools. These are described in the project plan, but updated descriptions are provided below in successive subsections of this section, focusing specifically on resource and tool formats (section 2.1 and 2.2) and whether there already are formal metadata descriptions available (section 2.3). Section 2.4 contains a brief discussion of the standardization requirements known so far for this first batch of META-NORD resources.

2.1. Language resources

Below, the language resources to be made available in the first META-NORD batch are listed.

2.1.1. Corpora (including treebanks), speech databases and multimodal resources

Language	Contact person(s)	Resource	Format
Danish	Bolette S. Pedersen <bspedersen@hum.ku.dk>, Lene Offersgaard <leneo@hum.ku.dk>	texts and text annotations	DK-CLARIN
	Bolette S. Pedersen <bspedersen@hum.ku.dk>, Lene Offersgaard <leneo@hum.ku.dk>	CLARIN LSP Corpora	Data format TEI P5 (Metadata format: DC, CMDI, TEI P5DK)
Estonian	Kaili Muurisep <kaili.muurisep@ut.ee>	Est comprehensive corpus	TEI
	Kaili Muurisep <kaili.muurisep@ut.ee>	treebank	TIGER-XML
Finnish	Hanna Westerlund <hmwester@cc.helsinki.fi>, Jussi Piitulainen <jussi.piitulainen@helsinki.fi>	finn tree-bank	CONLL-X format
	Hanna Westerlund <hmwester@cc.helsinki.fi>, Jussi Piitulainen <jussi.piitulainen@helsinki.fi>	Language bank of Finland	a local DTD close to an earlier version of TEI
Icelandic	Ásta Svavarsdóttir <asta@hi.is>	Icelandic speech corpus, 53 hours of transcribed Icelandic speech, synchronized text and sound files (text files are part of MÍM)	Sound files: .wav and mp3 Transcription files: TEI conformant xml format.
	Eiríkur Rögnvaldsson	Icelandic Parsed Historical	The file format is

Language	Contact person(s)	Resource	Format
	<eirikur@hi.is)>	Corpus (IcePaHC)	labeled bracketing text files with UTF-8 encoding. The format is compatible with any tool that operates on labeled bracketing and can be easily converted to different formats using existing or custom tools. A recommended search tool is CorpusSearch <http://corpussearch.sourceforge.net/> written by Beth Randall
	Eiríkur Rögnvaldsson <eirikur@hi.is)>	HJAL, Training material for a speech recognizer, collected and transcribed in 2003. Open source and free.	Sound files: .wav Transcription file: text.
	Eiríkur Rögnvaldsson <eirikur@hi.is)>	Pronunciation dictionary for Icelandic	Transcribed in IPA and SAMPA, Excel-file
	Kristín Bjarnadóttir <kristinb@hi.is>	BÍN, comprehensive full form database of modern Icelandic inflections, containing about 280,000 pardigms with over 5,8 million inflectional forms	proprietary XML
	Sigrún Helgadóttir <sigruhel@hi.is>	MÍM, 26 million word corpus of text and transcribed speech. the corpus is automatically tagged (available late 2011)	The corpus will be made available in TEI conformant xml format.
	Sigrún Helgadóttir <sigruhel@hi.is>	IFD, tagged Icelandic corpus with about 590 thousand words, tagging hand corrected	The corpus will be made available in TEI conformant xml format.
Latvian	Roberts Rozis <roberts.rozis@tilde.lv>	Latvian literature corpus	proprietary XML
	Roberts Rozis <roberts.rozis@tilde.lv>	Latvian-English legislation corpus	XCES
Norwegian	Gyri Smørdal Losnegaard <Gyri.Losnegaard@uib.no>		XML
Swedish	Lars Borin <lars.borin@svenska.gu.se>	Språkbanken's corpora	TEI P5

Table 2 Corpora (including treebanks), speech databases and multimodal resources

2.1.2. Lexical resources (including wordnets)

Language	Contact person(s)	Resource	Format
Danish	Bolette S. Pedersen <bspedersen@hum.ku.dk>, Lene Offersgaard <leneo@hum.ku.dk>	Wordnet	Data format: Princeton Wordnet format, Metadata: TEI P5
	Bolette S. Pedersen <bspedersen@hum.ku.dk>, Lene Offersgaard <leneo@hum.ku.dk>	STO Computational lexicon	XML and CSV
Estonian	Kaili Muurisep <kaili.muurisep@ut.ee>	wordnet	Princeton WordNet format
Finnish	Hanna Westerlund <hmwester@cc.helsinki.fi>, Jussi Piitulainen <jussi.piitulainen@helsinki.fi>	finn word-net	Princeton WordNet format
Latvian	Roberts Rozis <roberts.rozis@tilde.lv>	electronic dictionaries are stored in proprietary XML for dictionaries	proprietary XML
Lithuanian	Aurelija Tamulionienė <tamulioniai@gmail.com>	Dictionary of the Lithuanian language	Microsoft Access
	Aurelija Tamulionienė <tamulioniai@gmail.com>	Database of the Lexicon of Standard Lithuanian	MySQL
	Aurelija Tamulionienė <tamulioniai@gmail.com>	Geoinformational Database of Toponyms	PostgreSQL
	Aurelija Tamulionienė <tamulioniai@gmail.com>	Database of historical ethnic place names	MySQL
	Aurelija Tamulionienė <tamulioniai@gmail.com>	Database of Neologisms	MySQL
Swedish	Markus Forsberg <markus.forsberg@gu.se>	SB-LEX (linked lexical resources, including a framenet and a wordnet)	LMF

Table 3 Lexical resources (including wordnets)

2.1.3. Terminology resources

Language	Contact person(s)	Resource	Format
Icelandic	Ágústa Þorbergsdóttir <agustath@hi.is>	Icelandic term bank	TBX

Table 4 Terminology resources

2.2. Language tools

Language	Contact person(s)	Resource	Format
Finnish	Hanna Westerlund <hmwester@cc.helsinki.fi>, Jussi Piitulainen <jussi.piitulainen@helsinki.fi>	hfst	http://hfst.sourceforge.net/ xml (binaries and source code)
Icelandic	Hrafn Loftsson <hrafn@ru.is>	Apertium-is-en, a shallow transfer rule-based Icelandic to English machine translation system.Both programs and data are free and open source, available for download at <http://www.apertium.org>	XML format (http://wiki.apertium.org/ wiki/Monodix_basics)
	Hrafn Loftsson <hrafn@ru.is>	IceNLP, an open source Natural Language Processing (NLP) toolkit for analyzing and processing Icelandic text. The toolkit is implemented in Java and includes a tokeniser/sentence segmentiser, an unknown word guesser, a lemmatiser, a named entity recogniser, a linguistic rule-based tagger, a statistical tagger and a shallow parser. Available for download at <http://icenlp.sourceforge.net/>	Input/output in UTF-8 text format
	Hrafn Loftsson <hrafn@ru.is>	CombiTagger, an open source tool, implemented in Java, for developing and evaluating combined	Input/output in UTF-8 text format

Language	Contact person(s)	Resource	Format
		taggers according to a given combination method. Available for download at <http://sourceforge.net/projects/combitagger/>	
Swedish	Markus Forsberg <markus.forsberg@gu.se>	CLT Toolkit	Various
	Markus Forsberg <markus.forsberg@gu.se>	CLT Cloud	REST web services

Table 5 Language tools

2.3. Metadata

As stated in the introduction (Section 1), metadata for META-NORD resources should provide at least the following information in a standardized format suitable for machine harvesting:

(1) identification of the resource (including a persistent identifier), together with its;

(2) creation;

(3) annotation;

(4) provenance;

(5) documentation;

(6) usage;

(7) availability;

(8) licensing;

(9) distribution data. In addition, there should be;

(10) provenance information for the metadata items themselves.

In most cases, the resources and tools to be made available in the first META-NORD batch do not come equipped with this information, let alone in encoded as formal metadata. The main exceptions are these:

– Corpora in TEI or XCES format often have header elements containing at least some of this information, which can be automatically extracted.

– Some partners are already publishing structured metadata records for at least some of their resources: UCPH (DC, CMDI), UGOT (OLAC) and UHEL (the Language Bank of Finland is publishing OLAC – and the obligatory DC – through OAI-PMH for a number of corpora already).

2.4. Requirements for standardization

We can foresee that users will desire access to META-NORD language resources in at least the following three ways:

(1) *In toto*, i.e., the resource can be downloaded. This requires that the resource is in a standardized, well-documented format, or it won't be very useful to our target groups. It also requires that all IPR issues have been cleared.

(2) Online browsing either in a standard web browser or through a dedicated tool (e.g. a JWS application). Here, standardized metadata must provide sufficient information for a user to find the URL providing the application. However, the base resource may be in a proprietary format (although any export facility should provide a standardized format).

(3) In the form of a web service or other API. Here, standardized metadata is needed. Further, any data returned by a web service should be in a standard format.

Consequently, metadata and resource formats in META-NORD should support at least these three resource usage scenarios.

It should be evident from the tables in section 2.1 and 2.2 above that META-NORD resources and tools come in many formats. Some resources are in RDB formats (SQL, Access), some in proprietary formats, etc. For interoperability, such resources should probably be converted into other formats. As mentioned above in section 1.2, data format conversion is generally not a problem, and should be implemented in many cases, since partners may have invested heavily in such formats and in such cases we should simply consider a solution whereby conversion is made on demand into an interoperable export format. The only problem with this solution is that it will add complexity, since any change made to the original format must be accompanied by the corresponding change in the conversion utility.

With very few exceptions (some tools at UGOT), at least from the information provided, it seems that tools are not available with any form of standardized API usable in a distributed context.

According to the submitted information, many of the resources and tools lack an explicit and formal content model. This issue will need to be addressed in META-NORD.

Finally, only a small number of resources from a few partners have been provided with explicit metadata records in a CLARIN-compatible format.

3. Conclusion: Recommended metadata and resource formats for META-NORD

For the first batch of META-NORD resources and tools to be released at the end of project month 10 and in default of META-NET/ META-SHARE recommendations, and further considering the human and other resources available for the completion of this task, the following recommendations can be made on the basis of the information and discussion provided above:

3.1 Metadata formats and metadata

CMDI metadata fulfils all the META-NORD requirements and it is further a very likely candidate for adoption by META-NET/ META-SHARE.

Consequently, all META-NORD resources should be provided with CMDI records containing the following information:

(1) identification of the resource (including a persistent identifier), together with its

(2) creation;

(3) annotation;

(4) provenance;

(5) documentation;

(6) usage;

(7) availability;

(8) licensing;

(9) distribution data, as well as;

(10) provenance information for the metadata items themselves.

The content of the information fields should as much as possible come from standardized controlled vocabularies, e.g., ISO 639-1 or ISO 639-3 for language names, CLARIN licensing condition codes, etc.

The CMDI records should be stored in the format and manner required for automatic harvesting.

For metadata authoring skill transfer among META-NORD consortium partners, a project-wide metadata workshop should be arranged well before project month 10.

3.2 Data formats

In most cases, existing data formats can be kept – especially when they represent large investments in time and software solutions – and converters should be written.

For new resources and tools or for those where conversion of the base resource is desirable, the following formats are recommended:

– Corpora: TEI or (X)CES format (standoff annotation in ISO formats will be allowed);

– lexical resources: LMF or Princeton WordNet format;

– terminology resources: TBX;

– tools: at least as web services (if possible), described using WSDL.

3.3 Content models

It is recommended that META-NORD put a considerable effort into making content models of the partners' resources (and tools) as interoperable as possible. This can imply adopting more strictly structured formats, such as LMF rather than proprietary XML or SQL for lexical resources. Regardless of this, it will almost certainly imply a mapping to a set of standardized data categories, such as that of ISOcat. This can mean a considerable amount of work and careful consideration is needed in order not to waste effort. On the other hand, the rewards of the interoperability achieved in this way are potentially great.

In the same way as was proposed above regarding metadata authoring (section 3.1), META-NORD should arrange a project-wide workshop for knowledge transfer among partners about language resource content models and ISOcat, again well before project M10.

4. References

CLARIN deliverable D2.4. *Registry requirements. Metadata infrastructure for language resources and technology*. December 2008. <http://www-sk.let.uu.nl/u/D2R-4.pdf>

CLARIN deliverable D2R-5b. *Language resources and technology registry infrastructure* (version 2). January 2011. <http://www-sk.let.uu.nl/u/D2R-5b.pdf>

CLARIN deliverable D2.R6. *Requirement specification web services and workflow systems* (version 2). January 2010. <http://www-sk.let.uu.nl/u/D2R-6b.pdf>

ISO 12620 . *Computer applications in terminology – Data categories* .
First edition 1999-10-01.

ISO 24610-1 . *Language resource management – Feature structures – Part 1: Feature structure representation* . First edition 2006-04-15.

ISO 24613 . *Language resource management – Lexical markup framework (LMF)* .
First edition 2008-11-15 .

TEI P5. Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange (version P5). <http://www.tei-c.org/Guidelines/>

5. Tables