Standards for LRT

- a joint document by Marc Kemps-Snijders, Núria Bel, Peter Wittenburg, Daan Broeder, Dieter van Uytvanck (CLARIN), Laurent Romary (ISOTC37, TEI), Erhard Hinrichs (CLARIN) and Gerhard Budin (Flarenet) -
January 2009

This document is the basis for a joint web-site with recommendations for CLARIN. Each known name of a standard or best-practice guideline will be commented along a few criteria:

- the main function is indicated
- **Standard** will indicate whether it is a standard (++), a best practice in the field (+) or simply known (0)
- **State** will indicate what the state of it is: proven (++), ready (+) or in progress (0)
- **Pivot** will indicate whether the guideline is meant as a pivot mechanism (if so, indicated by +)
- **Advise** will indicated whether in CLARIN the usage should be obligatory (++), recommended (+) or whether CLARIN is neutral (0)
- further a comment will be given where necessary

In addition for easy overview we will use color coding supporting the (mostly three) options.

CLARIN should take care that all standards with a ++ advise will be supported in the infrastructure.

Name	Standard	State	Pivot	Advise	Function	Comment
General						
XML	++	++	+	++	text document structure description	CLARIN should require the usage of XML where feasible
W3C XML Schema	++	++		++	specification of classes of structures, i.e. constraining XML	CLARIN should require the existence of schemas when using XML
RNG (compact and XML variant)	++	++		++	same - but more simple to write	same (CLARIN does not state a preference)
RDF	++	++	+	++	mechanism to describe semantic relations	wherever possible an RDF output should be available
RDFS	++	++		+	specification of some semantics	certainly a recommended formalism
OWL	++	++		+	specification of semantics	certainly a recommended formalism
SKOS	++	++		+	more simple formalism to describe taxonomies	certainly a recommended formalism
URIs	++	++		+	General identifier system for resources on the Internet	ongoing debate whether URIs are stable
Handles	+	++		+	Persistent Identifier Framework for resources on the Internet	well-tested resolver system with additional services; CLARIN will offer a Handle issueing mechanism
URNs	++	0		0	URIs that do not specify an access protocol	yet no proven resolver available
Languages 639-3	++	+	+	++	unique specification of languages	new standard and still under debate, but a

Name	Standard	State	Pivot	Advise	Function	Comment
						requirement in CLARIN
Country codes (ISO 3166)	++	++		++	Country codes	Widely used as domain extensions
Script codes (ISO 15924)	++	++		++	Codes for the representation of names of scripts	
Protocols						
OAI PMH	++	++	+	++	a protocol for metadata harvesting	should be the preferable choice in CLARIN; for some difficult to implement
DCR API	0	0	+	+	an API to interact with the ISO DCR	should be offered to all DCR instances in CLARIN – a new version will soon be published at http://www.isocat.org/
WSDL	++	++	+	++	specification of web service API	should be the preferred option in CLARIN
SOAP	++	++	+	++	specification of data exchange in XML	should be the preferred option in CLARIN
REST	+	+		+	widely used simple web service API	no agreed specification language but widely used, so CLARIN may not ignore it
Terminology/Ont						
ISOcat/12620	++	+	+	++	model and software for the specification of linguistic concepts and terms	model is a standard; software is in progress; CLARIN will adopt this as a reference/pivot standard
DCR Profiles	++	0		++	concepts in ISOcat in different domains	CLARIN should strongly recommend the usage of DCR concepts or at least require to refer to them
EAGLES/ISLE	+	+		+	specification of linguistic concepts	since many of the defined concepts will be entries in ISOcat there is a natural follow up
GOLD	0	+		0	linguistic ontology	created in the Emeld project, there is much critique on the definitions
TBX	++	++		+	allows for the interchange of terminology data including detailed lexical information	should be a required standard in CLARIN for exchanging terminology data
TEI Tags	+	++		+	various tag sets defined by TEI (P5)	will be supported by CLARIN when elements are required
ISO 16642 TMF	++	++		+	Terminology Markup Framework	
Metadata						
Dublin Core DCMI	++	++	+	+	specification of 15 general metadata elements and a number of more detailed elements as qualified DC	should be generated as metadata delivered to all types of service providers such as DRIVER to support occasional users
OLAC	+	++	+	+	added refinements on DC elements	should be supported as a simple pivot format in LRT
IMDI	+	++		+	more detailed description set for various LR	is a widely used format and will be supported in CLARIN; elements will be in ISOcat
TEI Header Tags	+	++		+	specification of a wide number of elements	will be supported by CLARIN when elements are

Name	Standard	State	Pivot	Advise	Function	Comment
(module "header")					that can be used as metadata elements	required
CLARIN MDI	0	0	+	++	specification of a new component model that is making use of ISOcat element definitions	this will become the standard in CLARIN (when robustness has been proven)
METS	+	++		+	container format to exchange (meta-) data	will be recommended to be used as standard mechanism to package metadata and data for exchange purposes
MPEG21 DID	+	++		+	same	not that widely used as METS
MPEG7	+	++		0	for multimedia	stick to elements of text annotation
ORE	0	0		0	Collection description on the web	relatively new
MARC	+	++		0		widely used by libraries; it's a family of standards, one of which is MARCXML; stick to elements required for identifying potentially useful texts; note also that MARCXML is supported by METS
EAD	+	++		0		used by archives; stick to elements required for identifying potentially useful content
Media						
MPEG1/2/4	++	++		+	well-known media codecs and standards incl. compression	used for different purposes
H.264	++	++		+	state-of-the-art codec for MPEG4	currently the mostly used codec, also used for web streaming
mJPEG2000	++	++	+	+	new standard incl. lossless compression	currently the agreed standard for archiving
JPEG	++	++		+	standard for lossy image encoding	most widely used encoding scheme
PNG	++	++		+	free standard for lossless image encoding	Good alternative for TIFF
TIFF	++	++		+	family of image encoding schemes	not really standardized, used often with scanners
mp3	++	++		+	compressed audio codec	widely used for small devices
wav-linear PCM	+	++	+	+	direct digital format without compression	wav is a de facto standard and used for lin PCM encoding
General Text Formats						
HTML	++	++		+	mixed tag set for simple structuring and rendering	not a recommended format for structured information
PDF/A (= ISO 19005-1:2005)	+	++		+	widely used de facto standard for representing documents	not a recommended format for structured information
RTF	+	++		0	possible export format instead of DOC	not a recommended format, but supported
CSV					General text-based format often used to transfer tabular information	
LRT Text Formats						
LMF	++	+	+	+	lexicon format standardized by TEI -> ISO?	not yet widely used, CLARIN should use it as pivot format

Name	Standard	State	Pivot	Advise	Function	Comment
CES	+	?		?	corpus encoding format used for annotations	replaced by XCES
XCES	+	?		?	corpus encoding format used for annotations	based on XML, often used for annotated texts
TEI	+	++		+	well-designed textual structure	CLARIN will need to support TEI structured texts
CHAT	+	++		+	widely used format for child corpora	CLARIN will need to support CHAT
Shoebox/Toolbox	+	++		+	widely used format for field linguistics corpora	CLARIN will need to support SBX/TBX
Tipster	+	++		+	widely used format for annotated texts	CLARIN will need to support Tipster
EAF	+	++		+	widely used format for annotated media	CLARIN will need to support EAF
LAF	?	0		0	not yet clear whether this will emerge to a standard	
lexicography: ISO/DIS 1951	++	++		+	Presentation/representation of entries in dictionaries	
TMX	++	++		+	for parallel texts	
Text Encoding						
Unicode	++	++		++	General standard for text encoding	Supported encodings: UTF-8, UTF-16, UTF-32
ISO-*	++	++		+	General standard for text encoding	
ASCII (7/8 bits)	++	++		+	General standard for text encoding	

Revisions of this document
V4 (2009-02-03): addition of ISO country and script codes
V5 (2009-02-06): addition of TMX, TBF, ISO 1951 and the text encoding part
V6 (2009-03-03): addition of MARC, MPEG7 and EAD