ISLE Meta Data Elements for Session Descriptions Proposal

Version: 2.0 (first external version)
Date: 2 November 2000; MPI ISLE Team (isle@mpi.nl)

Introduction and motivation

This proposal draft for a scheme of meta-data elements is for the moment specifically directed towards describing multi-modal multi-media language corpora. We hope to extend the proposal in the near future with a special scheme for lexica.

We were guided by the desire to enable not only the resource discovery of major resources such as whole corpora but also be able to find individual resources from within corpora. For instance community members not only want to answer the question “find me all corpora with yaminjung speakers” but also “find me all sessions (recordings) with female yaminjung speakers younger then 60”. To be able to answer questions like this we cannot use an existing general meta-data scheme used for instance for library resource discovery such as Dublin-Core (with or without some extra qualifiers). We think we need a more extensive set of meta-data elements that captures the many needs of the different linguistic domains to easily find suitable resources.

Another guiding principle was the need to be able to browse the descriptions of language resources next to using them for automatic resource discovery. Although the two are similar, browsing capability requires “human readable” descriptions of (sub-) corpora and resources. Therefore you will find that the proposed set offers the possibility to specify these descriptions or link in (URL) references to other such “human-readable” descriptions at many levels.

You will notice that the meta-transcriptions only contain references to real language resources such as audio/video files and transcriptions and annotations. All these references are accompanied by a structure specifying access restrictions for these resources . In our concept the access to meta-data in the meta-transcriptions is always free although the meta-data referring to individual persons may be rendered anonymous. The access to the resources themselves though may be restricted.

Flexibility: the possibility to have sub-communities add their own specific descriptions is approached in two ways. At different levels of the session description it is possible to add a list of keys in the form of name/value pairs. This possibility can be exploited by having sub-communities defining their own sets of required keys. Secondly the meta-description is characterised by a meta-description format identification. This identification will tell tools working with meta-descriptions what they can expect with respect to the structure of the meta-descriptions and the set of meta-data elements used. The format identification could also be used to inform specifically tailored tools to look for specific extensions to the basic scheme and act accordingly. This functionality is closely connected to the way the meta-data elements will be implemented and will pose extra requirements regarding this implementation. For the moment it seems wise to avoid the matter of structure and implementation and concentrate on discussing the appropriateness and sufficiency of the proposed meta-data element set for our purposes.

The sheer number of proposed elements may let people believe that it is a heavy burden to have to supply all this information. It should be taken into account that in most projects the meta-descriptions for different sessions vary only in a few fields. There will be special tools that allow users to use existing meta-transcriptions to generate new ones. This will considerably reduce the amount of typing involved.

Finally we need to say something on the set of meta-data elements that should be minimaly specified. Evidently not all the information that can be specified with the proposed set of meta-data elements is always available. This is specifically the case for legacy resources or very specialistic resource. Therefore only those elements should be mandatory that are needed for the correct functioning of tools working with the meta-descriptions, these are few.

1 Meta Data Element Proposal Overview

Meta Transcript				
			* Date	Meta-transcript creation date (DC:Date)
			* Version	Meta-transcript version of the session
			* Format ID	Meta-transcript description format identifier
			* Generated	Indicates how the Meta-transcript is produced { Automatic / hand / hand checked }
			History	Link to the change history
	Session			
			* Name	Short name of the session
			Title	Title of the session (DC:Title)
			Description	Description of the session (DC:Description)
			Date	Date of the recording (DC:Date)
			Continent	Continent of the recording
			Country	Country of the recording
			Region	Region of the recording
			Address	Address of the recording
		Project		
			* Name	Short name or abbreviation of the project
			Title	Project title(DC:Title)
			Description	Project description (DC:Description)
			Id	Identifier for the project (DC:Identifier)
			Contact	Name and address of the contact for the project

				Institute / affiliation	Name / address of affiliated institute
		Creator			
				Name	Name of the session creator (DC:Creator)
				Address	Address of the session creator (usually principal interviewer)
				Link/URL	Link to external information on the creator of the session
		Content			
				Description	Elaborate description of the content (DC:Description)
				Genre	Genre of the content { Reportage / Dialogue / Singing / Novel / … } (DC:Type)
				Modalities	Modalities of the content { Speech / Writing / Gestures / Facial Expressions / Pointing Gestures / … }
				Keys	A list of attribute-value pairs used to describe the specifics of the content
			Languages used		
				Description	Description of the languages used in the session (DC:Description of languages)
				Main language	Indicates the main language used in the session
				Language+	All languages used in the session
		Participants			
			Interviewer +		
				Name	Name of the interviewer used in the transcription
				Full name	Full name of the interviewer
				Code	Short unique code to identify the interviewer
				Role	The role of the interviewer
				Language	The first language of the interviewer
				Ethnic group	The interviewer's ethnic group
				Sex	The interviewer's sex
				Link/URL	A link to the interviewer (DC:Identifier)
			Informant +		
				Name	Name of the informant used in the transcription
				Full name	Full name of the informant
				Code	Short unique code to identify the informant
				Role	The role of the informant

					First Language	The first language of the informant
					Language +	Informant's languages
					Ethnic group	Informant's ethnic group
					Sex	Informant's sex
					Age/Born	Informant's age indication
					Education	Informant's education
					Anonymous	Used when the name of the informant is replaced by a pseudo name to make them anonymous { True / False }
					Keys	A list of attribute-value pairs to describe specific characteristics of the informant
				Contributary +		
					Name	Name of an additional contributor used in the transcription
					Code	Short unique code to identify an additional contributor in the transcript
					Role	Role of the additional contributor in the transcript
			Resources			
				Media File +		
					Access	Access rights of the media file
					URL	URL to the media file (more narrow than DC:Identifier)
					Size	Size of the media file
					Type	Type of media file { Audio / Video / Photo }
					Format	Format of the media file { … }
					Quality	Quality of the media file { Low / Medium / High }
					Position	Start / end time
					Tool +	List of tools for the resource
				Transcription / Annotation File +		
					Access	Access rights of the transcription / annotation file
					URL	Link to transcription / annotation file (more narrow than DC:Identifier)
					Size	Size of the transcription / annotation file
					Type +	Type of the transcription / annotation file { Orthography / Phonetics / Phonemics / … }
					Format	Format of the transcription / annotation file { CHAT / Shoebox / … }

				Font / encoding table	Font used in the transcription file or encoding scheme name / reference used in annotations
				Date	Date of transcription/annotation
				Lexicon name & reference	Link to a lexicon
				Anonymous	Used when the names in the transcripts are replaced by pseudo names to make them anonymous { True / False }
				Tool +	List of tools for the resource
			Anonymous		
				Access	Access rights of the name conversion file
				URL	URL to the name of the conversion file that makes it possible to convert pseudo names into real names (DC:Identifier)
			Media Carrier +		
				Access	Access rights of the media carrier
				Storage format	Physical storage format of the media { CD / MD / DAT / … }
				Quality	Quality of the media carrier
				Position	Index, Start / end time on the media carrier
				Id	Identification of the media carrier (DC:Identifier)
		References			
				Publications	Link to publications associated with the content

Bold & italic indicates the attribute is a structured sub element
A '+' sign indicates a list
A '*' sign indicated an element is mandatory

At every place where a description field is specified it is possible to augment or replace this description with a (list of) URL reference(s) to a "human-readable" resource for instance a web-page containing descriptive information.

2 Meta Data Structured Sub-Elements

Keys	A list of name value pairs
Name1 = Value 1	Associate Value 1 with domain specific element Name1
Name2 = Value 2	Associate Value 2 with domain specific element Name2
“	“
“	“

Language	Language elements
Id	ISO 639-2 identifier
Name	Readable/understandable name
Description	Elaborate description
Link	Link to language DB

Access	Access rights
Availability	Availability of the resource [Free / restricted / none]
Description	Restriction description
Date	Date of access rights evaluation
Owner	Name / address if applicable
Publisher	Name / address of publisher if any
Contact	Address, e-mail of organisation to obtain access

Contact	Name and address of the contact for the project
Name	Name of contact
Address	Address of contact
E-mail	E-mail address of contact
Organization	Organization of contact

3 Meta Data Element Definitions

Meta Transcript Contains information about the meta description file itself. All of the elements are generated automatically when a meta-description tool is used. These elements serve administrative purposes and are used by tools that work with meta descriptions.

Meta Transcript . Date Indicates the date of when the meta description file is created. For example, when an meta editor is used to create a new meta description file, it should save the date of creation in this element.

Meta Transcript . Version The version of the content of the meta description file. When metadata in the meta description file is changed, this version number should be incremented.

Meta Transcript . Format ID The format identifier of the meta description file. This is used to indicate which meta description format and revision is used to describe the meta elements.

Meta Transcript . Generated Indicates how the meta description file is produced. A meta description file can be generated by a certain tool, by hand or checked by hand after its generated [Automatic / hand / hand checked].

Meta Transcript . History Link to the change history of the metadata in the meta description file.

Session This concept bundles all information and resources belonging to the record of a linguistic action and its description. If an interviewer questions an informant the resulting session does not only contain the recording of that interview but also the transcription and annotations and also for instance any photo images that were taken of this interview. It may well be that a researcher decides that one interview contains in fact more then one session if for instance the informant is asked to perform different tasks during that interview. This is all at the discretion of the researcher. The session is just a concept that can be used to create order when dealing with many linguistic resources.

Session . Name A short identifier for a particular session, typically a code name of one or two words or strings. This identifier distinguishes the session from others in the same (sub-)corpus.

Session . Title A title for the Session. Something like “Frog story”. The semantics of this element maps with DC:Title.

Session . Description An elaborate description of this session with information pertaining to the Session level. A description of the content is better specified at the level of the Content . description element.

Session . Date The date when the primary data (in general audio or video data) of the session was recorded.

Session . Continent Continent locator of place where the session was recorded or originated. This element enables linking to the world map of languages. Limited set.

Suggested possible values are { Europe, Asia, Australia, Oceania, North America, Middle America, South America, Africa}.

Session . Country Country where the session was recorded or originated.

Session . Region Region of the Session . Country where the session was recorded or originated.

Session . Address Address where the session was recorded or originated. For instance if recording sessions took place at an institution, the address of the institute is meant.

Project If the sessions were made within the context of a project, the project element contains information regarding this project.

Project . Name A short name or abbreviation of the project.

Project . Title Project title. Comparable to DC:Title.

Project . Description Elaborate description of the project.

Project . Id A short identifier for the project. Comparable to DC:Identifier.

Project . Institute / affiliation Gives the name and address of affiliated institute associated with the project.

Creator Groups information about the creator of the session.

Creator . Name The name of the person who is responsible for the creation of the session (DC:Creator).

Creator . Address The address of the institution affiliated with the person responsible for the creation of the session.

Creator . Link URL link to external information of the institution affiliated with the person responsible for the creation of the session.

Content Groups information about the content of the session.

Content . Description An elaborate description of the content of the session. Comparable to DC:Description.

Content . Genre Indicates the genre of the content of the session. A limited set of values is allowed for genre { Reportage / Dialogue / Singing / Novel / … }. Comparable to DC:Type.

Content . Modalities Gives a list of modalities of the content of the session. A limited set of values is allowed for modalities { Speech / Writing / Gestures / Facial Expressions / Pointing Gestures / Facial expression / Pointing gesture / … }.

Content . ***Keys*** A list of attribute-value pairs used to describe the domain specific characteristics of the content

Languages used Groups information about all the languages used in the session.

Languages used . Description A description of the languages used in the session.

Languages used . Main language Indicates the main language used in the session. This is the language which is the main interest of the interviewers.

Languages used . ***Language +*** A list of all the languages used in the session

Participants Groups information about all the participants in the session.

Interviewer Groups information about an interviewer in the session.

Interviewer . Name The name of the interviewer as it is used by others in the transcription.

Interviewer . Full name Full name of the interviewer.

Interviewer . Code Short unique code to identify the interviewer. The code is used in the transcription and annotations to identify parts belonging to this specific participant.

Interviewer . Role The role of the interviewer, usually just “interviewer”, but the interviewer might play a separate role within the community.

Interviewer . Language The first language of the interviewer.

Interviewer . Ethnic group The interviewer’s ethnic group.

Interviewer . Sex The interviewer’s sex.

Informant Groups information about an informant in the session.

Informant . Name Name of the informant as it is used in the transcription.

Informant . Full name Full name of the informant.

Informant . Code Short unique code to identify the informant. It is used in the transcription and annotations to identify parts belonging to this specific participant.

Informant . Role The role of the informant. For instance when interviewing part of a family group, “Role” should specify the mutual relations within the group.

Informant . First Language Indicates the first language of the informant.

Informant . ***Language +*** A list of structured ***language*** elements containing the Informant’s languages.

Informant . Ethnic group Informant’s ethnic group.

Informant . Sex Informant’s sex.

Informant . Age/Born Indication informant's age at the time of the session recording or date of birth.

Informant . Education Informant's education.

Informant . Anonymous Used when the name of the informant is replaced by a pseudo name to make them anonymous. If anonymous is set to 'True', the real name of the informant can be obtained from the 'conversion file' when access is granted.

Informant . ***Keys*** A list of attribute-value pairs to describe domain specific characteristics of the informant.

Contributary Groups Information about an additional contributary of the session.

Contributary . Name Name of an additional contributary as it is used by others to identify that person in the transcription.

Contributary . Code Short unique code to identify an additional contributary. It is used in the transcription and annotations to identify parts belonging to this specific participant.

Contributary . Role Role of the additional contributary.

Resources Groups information about all the resources associated with the session.

Media File Groups information about the media file.

Media File . URL URL to the media file.

Media File . Size Size of the media file.

Media File . Type Type of media file { Audio / Video / Photo/ … }.

Media File . Format Format of the media file { … }.

Media File . Quality Quality of the media file { Low / Medium / High }.

Media File . Position Start / end time.

Transcription / Annotation File Information about the transcription or annotation file.

Transcription / Annotation File **.** URL Link to transcription / annotation file.

Transcription / Annotation File . Size Size of the transcription / annotation file.

Transcription / Annotation File . Type + Type of the transcription / annotation file { Orthography / Phonetics / Phonemics / Morpho-syntax / Glossing / … }.

Transcription / Annotation File . Format Format of the transcription / annotation file { CHAT / Shoebox / … }.

Transcription / Annotation File . Font / encoding table Font used in the transcription file or encoding scheme name/reference used in annotations.

Transcription / Annotation File . Date Date of transcription/annotation.

Transcription / Annotation File . Lexicon name & reference Link to a lexicon.

Transcription / Annotation File . Anonymous Used when the names in the transcripts are replaced by pseudo names to make them anonymous { True / False }.

Conversion File Groups information about the name conversion file for persons who are anonymous in the transcript.

Conversion File . URL URL to the conversion file used to convert the pseudo names into real names.

Media Carrier Groups information about the media carrier.

Media Carrier . Storage format Physical storage format of the media. Values from a limited set are allowed { CD / MD / DAT / … }. Comparable to DC:Format.

Media Carrier . Quality Quality of the recorded data on the media carrier.

Media Carrier . Position Index, Start / end time on the media carrier.

Media Carrier . Id Short code to identify the media carrier.

References Groups information about references associated with the session.

References . Publications Link to a list with publications associated with the session.

Keys Groups domain specific attribute-value keys.

Keys . Name = Value Associate Value with domain specific element Name.

Language Groups information about a language.

Language . Id ISO language identifier.

Language .Name Readable/understandable name.

Language .Description Elaborate description.

Language .Link Link to language DB.

Access Groups information about access rights.

Access . Availability Availability of the resource { Free / Restricted / None }.

Access . Description Restriction description.

Access . Date Date of access rights evaluation.

Access . Owner Name / address if applicable.

Access . Publisher Name / address of publisher if any.

Access . ***Contact*** Address, e-mail of organisation to obtain access.

Contact Groups information about a contact person.

Contact . Name Name of contact person.

Contact . Address Address of contact person.

Contact . E-mail E-mail address of contact person.

Contact . Organization Organization of contact person.