CAVA (human Communication: an Audio Visual Archive)

Project Report

Descriptive metadata for the CAVA repository

Matt Mahon
CAVA Project Officer
October 2009

http://www.ucl.ac.uk/ls/cava

CAVA Project: Metadata
Matt Mahon, October 2009
Contact: lib-cava@ucl.ac.uk

DESCRIPTIVE METADATA FOR THE CAVA REPOSITORY

CONTENTS

INTRODUCTION

This document describes the metadata schema devised by the CAVA Project Teamfor use in the CAVA repository.

In addition to collecting and standardising the quality of the data, the CAVA project also aims to make easy discovery of the data possible. CAVA uses an in-house descriptive metadata schema based on the ISLE MetaData Initiative (IMDI), a standard designed for language resources.

The nature of the data presents some crucial challenges to the creation of metadata. Implementing the full IMDI standard would be too time-consuming and costly, for both the project team and depositors. IMDI offered the best metadata standard to start from, as that initiative was already concerned with describing multi-media and multi-modal language resources. The UCL Deafness, Cognition and Language Research Centre (DCAL) subset offered an IMDI-based schema that described actor conditions such as deafness, but this was too specific for CAVA needs. The schema described here represents a pragmatic solution which incorporates the DCAL subset into a more general description of other conditions and multiple actors.

The metadata schema and indicative vocabularies were drafted by the CAVA project team and tested against a small pilot dataset. These two products were then assessed by a wider user group consisting of the members of the UCL Centre for Applied Interaction Research (CAIR), whose membership makes up the initial wave of depositors to the repository, and the CAVA project Steering Group.

CAVA METADATA SCHEMA

Table 1 shows the CAVA metadata. Elements marked with (c) are encoded using a controlled vocabulary. Elements marked with [brackets] may be left blank.

TABLE 1.

No.	*Object +*		
1	Identifier		
2	Date (c)		
3	Original format (c)		
4	Format history		
	Location (sub)		
5		Country (c)	
6		Description	
	Project +		
7		Name	
8		ID	
		Contact (sub)	
9			Name
10			Contact's organisation
11		Longitudinal project (boolean)	
12		Description	
	Content +		
13		Genre	
14		Subgenre	
15		Communication Context	
		Languages (sub)	
16			Number of languages (c)
17			Spoken language ID (c)
18			Sign language ID (c)
19			Language variety
20			Communication modes
		Transcription (sub)	
21			Transcription (boolean)
22			[Transcription format]
	Actors +		
23		ID	
24		Age (c)	
25		Sex (c)	
26		[Occupation or previous occupation]	
27		[Actor notes]	
		Condition (sub)	

28			Condition
29			Condition subtype
30			Cause of condition
31			Onset of condition
32			Intervention history
33			Family history
34			[Hearing status]
35			[Vision status]
36			[Handedness]
37			[Sign language experience]
		Education (sub)	
38			[Education leaving age] (c)
39			[School Type]
40			[Class Kind]
41			[Education Model]
42			[Boarding School] (boolean)
43		Secondary actor(s) notes	
	Access +		
44		Rights (c)	
45		Rights evaluation date (c)	
46		Owner	

ELEMENT DESCRIPTIONS AND INDICATIVE VOCABULARIES

Table 2 below shows the element descriptions and indicative vocabularies for the CAVA metadata. It works as follows:

1.	*ELEMENT*	*DESCRIPTION*
		INDICATIVE VOCABULARY

All vocabulary lists here are open and may be added to, although the use of given vocabulary is highly encouraged. Multiple entries in each element are also recommended wherever appropriate, separated with commas.

TABLE 2.

OBJECT +		
1.	Identifier	The name of the session (file).
		*Controlled – see **Table 3**.*
2.	Date (c)	The date the file was created. YYYY-MM, or circa.
		Controlled
3.	Original format (c)	The format in which the recording was first made.
		Controlled
4.	Format history	An open description of any changes to the format of the recording.
		Free text. For example, "Converted to AVI, MPEG-1 and WAV for deposit"
Location (sub)		
5.	Country	The country in which the recording was made.
		Controlled
6.	Description	An open description of the location.
		Name the town or city and more specific location. For example, if Country is 'United Kingdom', the description might include "London, Primary Care Trust clinic". **It is not appropriate to name the institution where the recording took place if this may help to identify the participants.**
PROJECT+		
7.	Name	The name of the project for which the recording was made.
		Free text. For example, "EAL deaf children"
8.	ID	The ID number of the project.

		Alphanumeric. For example, "HMM-DOH" or "ESRC R000239306"
Contact (sub)		
9.	Contact name	The name of the primary researcher(s) on the project. Free text. For example, "Dr Suzanne Beeke"
10.	Contact's organisation	The organisation at which the primary researcher(s) are based. Free text.
11.	Longitudinal project (boolean)	Is this session part of a longitudinal dataset? { yes | no }
12.	Project description	An open description of the project. Free text.
CONTENT+		
13.	Genre	The genre of the session. The following open vocabulary is suggested: • Alone • Group • One:One
14.	Subgenre	The subgenre of the session. The following open vocabulary is suggested: • Adult and adult • Adult and speech and language therapist • Adult parent and adult child • Child and child • Child and parent • Child and sibling • Child and teacher • Child and speech and language therapist • Family group • Partners • Peer group • Spouses
15.	Communication context	The communication context. The following open vocabulary is suggested:

		• Assessment session • Booksharing • Free play • Institutional conversation • Peer conversation • Teaching session • Therapy session
Languages (sub)		
16.	Number of languages (c)	The number of languages, spoken or signed, used in the recording.
		Controlled
17.	Spoken language ID (c)	The ID of the spoken language(s) used.
		Controlled
18.	Sign language ID (c)	The ID of the sign language(s) used.
		Controlled
19.	Language variety	The variety of languages used.
		List any dialect or further language detail which is not recorded by the encoding for language IDs. For example, if Spoken language ID is ‘eng’, Language variety may include ‘Estuary’ or ‘Wife using Malay English, husband responding in Tamil’ and so on.
20.	Communication modes	Communication modes used.
		An open description of modalities used in the recording. The following open vocabulary is suggested: • Augmentative/alternative communication aid • Cultural gestures • Deictic (pointing) gestures • Emotional states • Enactment • Eye gaze • Haptics (touch) • Signs (from Sign Language lexicon) • Speech • Writing • Drawing
Transcription (sub)		

21.	Transcription (boolean)	Are there any transcripts associated with the session?
		{ yes | no }.
22.	[Transcription format]	An open description of the type of transcription documents associated with the session.
		Use the list below, or name the appropriate file extension or FourCC from the controlled vocabulary 'Original Format'. The following open vocabulary is recommended: • Unknown • Unspecified • Atlas TI • ELAN • Rich Text Format • Transana
ACTOR+		
23.	ID	Unique identifier for the primary actor in the session.
		Alphanumeric. This should correspond to the owner's encoding as used in any associated transcriptions. **It is not appropriate to name the actor.**
24.	Age (c)	The age of the primary actor.
		Controlled
25.	Sex (c)	The sex of the primary actor.
		The following open vocabulary is used: • Unknown • Unspecified • Male • Female • Transsexual
26.	[Occupation or previous occupation]	The occupation or previous occupation of the primary actor.
		Free text. Leave blank if the actor is a child.
27.	[Actor notes]	Any further notes on the actor.
		Free text.
Condition (sub)		
28.	Condition	The general condition of the primary actor.
		The following open vocabulary is used: • Unknown

		• Unspecified • Age related hearing loss • Aphasia • Autistic spectrum disorder (Adult) • Autistic spectrum disorder (Child) • Cerebral Palsy • Cognitive communication disorder • Deafness (Adult) • Deafness (Child) • Dementia • Dysarthria • Dyslexia • Dyspraxia • Language impairment (Child) • Language Impairment (Adult) • Learning Disability (Adult) • Learning Disability (Child) • Other physical disability • Progressive neurological • Second/additional language • Stammering • Typically ageing • Typically developing
29.	Condition subtype	An open description of the specific condition of the actor.
		More detail on the actor's condition. For example, if the condition is 'Deafness (Child)', then the Subtype may be 'Sensori-neural bilateral hearing loss'; if the condition is 'Aphasia' then the Subtype may be 'Agrammatic aphasia' etc. The following open vocabulary is suggested: • Unknown • Unspecified • [free text]
30.	Cause of condition	The cause of the condition.
		The following open vocabulary is suggested: • Unknown • Unspecified

		• Congenital • Stroke • Head injury • Brain tumour
31.	Onset of condition	An open description of the onset of the condition.
		If dates are included, please format as ‘YYYY-MM’ or ‘YYYY-MM-DD’. The following open vocabulary is suggested: • Unknown • Unspecified • [free text]
32.	Intervention history	An open description of the history of interventions.
		An open description of the history of interventions. If dates are included, please format as ‘YYYY-MM’ or ‘YYYY-MM-DD’. The following open vocabulary is suggested: • Unknown • Unspecified • “YYYY-MM, [intervention]; YYYY-MM, [intervention]”
33.	Family history	An open description of the history of the specific condition in the actor's family.
		A description of the history of the condition in the actor’s family. The following open vocabulary is suggested: • Unknown • Unspecified • [free text]
34.	[Hearing status]	The hearing status of the primary actor
		The following open vocabulary is suggested: • Unknown • Unspecified • Deaf • Hard-of-hearing • Hearing • No reported difficulties
35.	[Vision status]	The vision status of the primary actor.
		The following open vocabulary is suggested: • Unknown

		• Unspecified • Blind • Glasses for reading • Partially sighted • No reported difficulties
36.	[Handedness]	The handedness of the primary actor.
		The following open vocabulary is suggested: • Unknown • Unspecified • Ambidextrous • Left • Right
37.	[Sign language experience]	An open description of the actor's exposure to sign language.
		An open description of the actor's exposure to sign language. Give dates in the form 'Years; months', or 'birth'.
Education (sub)		
38.	[Education leaving age]	The age at which the (adult) actor left school.
		Controlled
39.	[School type]	The type of school the primary actor attends/attended.
		The following open vocabulary is suggested: • Bilingual (speech-sign) home programme • College • Home schooling • Preschool/nursery • Primary school • Secondary school • Special school • University • Vocational training
40.	[Class kind]	The type of class the primary actor attends/attended.
		The following open vocabulary is suggested: • Class in mainstream school • Class in special school • Individually integrated in mainstream class

		• Mainstream class
41.	Education model]	The education model employed in the class.
		The following open vocabulary is suggested: • Bilingual (spoken) • Bilingual/bimodal (speech and sign) • Oral with sing language interpreter • Oral/natural language • Sign only
42.	[Boarding school] (boolean)	Was/is the school a boarding school?
		{ yes | no }
43.	Secondary actor(s) notes	Any notes on secondary actors - their ID, roles etc.
		Free text. **It is not appropriate to name any secondary actors.**
ACCESS+		
44.	Rights (c)	The tier of access to which this session belongs.
		Controlled
45.	Rights evaluation date (c)	The date of access rights evaluation. YYYY-MM-DD.
		Controlled
46.	Owner	The owner of the resource. May be the same as The owner of the resource. May be the same as Project . Contact . Name, or may be an institution.
		Free text. May be the same as Contact Name, or may be an institution.

ENCODING SCHEMES:

Table 3 shows how elements which conform to particular external standards should be completed. Please follow the links provided to see full details of each external scheme.

TABLE 3.

1.	Identifier:	The identifier of each recording is controlled according to the owner’s own encoding. This must correspond with the name of the file as deposited.
2.	Date (c):	Dates are encoded in YYYY-MM or YYYY-MM-DD format, according to a profile of **[ISO8601]** as described in **[W3CDTF]**.

3.	Original format (c):	If the format is analogue, please name it in free text, for example “VHS” or “Audio cassette”. If the file is born digital, give a file extensions or FourCC codes, for example AVI, WAV, MPEG-1 etc. These are encoded by **Filext.**
5.	Country:	The country is encoded according to **[ISO3166-1]** 2- or 3-digit codes or in the longhand specified by the ISO code.
16.	Number of languages (c):	An integer.
17.	Spoken language ID (c):	Spoken language ID can be encoded according the following two schemas. If a language used does not appear on these lists, please name it in the Language variety field. • **[ISO639-1]**, which specifies the code set for language identification in the form of a two-letter code, or **[ISO639-2]** which specifies the code set for language identification in the form of a three-letter code. • The three-letter codes from the **[ETHNOLOGUE]** list from SIL International are allowed by using the prefix 'x-sil-' for the three-letter code (See **[LANGID]** for more information). For example, one could enter the language identifier 'x-sil-dut' to indicate the Dutch language.
18.	Sign language ID (c):	Sign language ID is encoded according to**[ISO639-2],** which specifies the code set for language identification in the form of a three-letter code. See **[SIGNWRITING]** for a mapping of signed languages to the ISO standard.
24.	Age (c):	Age is encoded as ‘years;months’, as specified by Codes for the Human Analysis of Transcripts **[AGECHAT]**.
38.	[Education leaving age] (c):	Age is encoded as ‘years;months’, as specified by Codes for the Human Analysis of Transcripts **[AGECHAT]**.
44.	Rights (c):	**TO BE SPECIFIED.**
45.	Rights evaluation date (c):	The date is encoded according to a profile of **[ISO8601]** as described in **[W3CDTF]** and follows the YYYY-MM format.