



ISLE Metadata Initiative (IMDI)

PART 1 B

**Metadata Elements  
for  
Catalogue Descriptions**

Draft Proposal Version 2.1

June, 2001

# INDEX

1	INTRODUCTION.....	3
2	CATALOGUE ELEMENTS OVERVIEW.....	4
3	METADATA ELEMENT DEFINITIONS.....	6
3.1	CATALOGUE.....	6
3.1.1	<i>Catalogue . Name</i> .....	6
3.1.2	<i>Catalogue . Title</i> .....	7
3.1.3	<i>Catalogue . Id</i> .....	7
3.1.4	<i>Catalogue . Description</i> .....	7
3.1.5	<i>Catalogue . Subject Language</i> .....	7
3.1.6	<i>Catalogue . Document Language</i> .....	7
3.1.7	<i>Catalogue . Location</i> .....	7
3.1.8	<i>Catalogue . Content Type</i> .....	8
3.1.9	<i>Catalogue . Format</i> .....	8
3.1.10	<i>Catalogue . Quality</i> .....	8
3.1.11	<i>Catalogue . Smallest Annotation Unit</i> .....	9
3.1.12	<i>Catalogue . Application</i> .....	9
3.1.13	<i>Catalogue . Date</i> .....	9
3.1.14	<i>Catalogue . Project</i> .....	9
3.1.15	<i>Catalogue . Publisher</i> .....	9
3.1.16	<i>Catalogue . Authors</i> .....	9
3.1.17	<i>Catalogue . Size</i> .....	9
3.1.18	<i>Catalogue . Distribution Form</i> .....	10
3.1.19	<i>Catalogue . Access</i> .....	10
3.1.20	<i>Catalogue . Pricing</i> .....	10
3.2	SUB-SCHEMAS.....	10
3.2.1	<i>Access</i> .....	10
4	VOCABULARIES.....	11
4.1	CATALOGUE . CONTENT TYPE.....	11
4.2	CATALOGUE . SMALLEST ANNOTATION UNIT.....	11
4.3	CATALOGUE . APPLICATION.....	11
5	ENCODING FORMATS.....	12
5.1	LANGUAGE IDENTIFIER ENCODING.....	12
6	REFERENCES.....	13
	APPENDIX A : CATALOGUE METADATA INVENTORY.....	14
	APPENDIX B : REVISION HISTORY.....	15

# 1 Introduction

In discussing the current set of IMDI metadata elements with various people we came to the conclusion that some metadata elements that describe “published” corpora at the top level (for instance the number of CD’s) are not present in the current IMDI set or are inappropriate to describe at the Session level. We therefore tried to make an inventory of such elements using existing description formalisms used by institutions that deal with “published corpora” such as [\[ELRA\]](#) and [\[LDC\]](#). We call the set of metadata elements that describe “published corpora” at the top-level “catalogue” metadata elements for language resources.

This proposal has been put forward after studying the current catalogue structure of ELRA, LDC and the UHLCS metadata requirements (compiled by Pirkko Suihkonen). It takes as a starting point all the fields specified by ELRA / LDC and looks for any matching UHLCS fields (see Appendix A).

What we have done is to categorise the fields and try to determine which of them are already covered in the IMDI proposal for SESSIONS. Also we try to determine which elements of the IMDI proposal for SESSIONS would better fit in this catalogue metadata set or would make sense to be duplicated in the catalogue set.

It would be feasible to implement this catalogue metadata set for language resources in Dublin-Core [\[DCMES\]](#). This is to be expected since it is metadata about published material without much information on the constituting parts. Therefore we have made a suggestion for DC equivalents with all the proposed catalogue elements. Sometimes it appears that a choice is possible, this has been indicated by a question mark.

Some concluding remarks:

- We are not yet sure that a separate IMDI catalogue metadata set is useful in all cases. It would seem so if you were looking for a corpus you can obtain as a unit and which has probably been produced for your needs (e.g. language technology applications).
- The flat description of a whole corpus by this set seems to make it ideal for easy identification of corpora by less-specific (in comparison to the IMDI Session vocabulary) search such as OAI.

All proposed DC qualifiers are our own except for Subject.language which is [\[OLAC\]](#)’s.

## 2 Catalogue Elements Overview

Element Name		DC Equivalent	Definition	Encoding	Connection to Session Elements
Name		Title.short	The name of the corpus	string	duplicate
Title		Title	The title of the corpus	string	duplicate
Id +		ID	A Unique identifier for the corpus. For example an ISBN.	string	None
Description +		Description	A description of the corpus	string	duplicate
Subject Language		Subject. language ( <a href="#">[OLAC]</a> qualifier)	The language subject of analysis	<a href="#">language identifier</a>	Summation of
Document Language		Language	The language the document is in	<a href="#">language identifier</a>	Summation of
Location		Coverage	Groups the information about the location of where the corpus content was made	continent, country, region	Generalisation
	Continent		The continent of where the corpus content was made	closed controlled vocabulary	
	Country		The country of where the corpus content was made	<a href="#">ISO3166-1</a>	
	Region		The region or sub-region of where the corpus content was made	string	
Content type		Type? Subject.type?	The type of the corpus	open vocabulary  Type = { <i>written, speech, terminology</i> } Subtype = { <i>Corpus, Monoling. Lex., Multiling. Lex., Telephone speech, Desktop/microphone, Multimodal/Multimedia, Other speech related</i> } <b>See ELRA</b>	Generalisation
Format		Format	Groups information about the formats used in the corpus	text, audio, video	Implicit in resource description
	Text		The format of the text used in the corpus	Text (encoding, ...)	
	Audio		The format of the audio used in the corpus	Speech ( sample freq, stereo/mono sample width,...)	
	Video		The format of the video used in the corpus	Video( MPEG1, Quicktime, ...)	
Quality		Format.quality	Groups information about the quality of the corpus content	audio, video	

	Audio		The quality of the audio data in the corpus	closed controlled vocabulary {1..5}	
	Video		The quality of the video data in the corpus	closed controlled vocabulary {1..5}	
Smallest Annotation Unit	Format.unit		The smallest unit of annotation used in the corpus	open vocabulary {paragraph, utterance, word, phoneme, ...}	In future version part of annotation unit
Application	Type. application ? Subject. application?		The application domain of the corpus	open vocabulary list (LDC set seems pretty exhaustive)	Summation of
Date	Date.Issued		Publishing date	YYYY-MM-DD	none
Project	Creator. project			string	duplicate
Publisher	Publisher		An entity responsible for making the resource available	string	none
Authors	Creator		An entity primarily responsible for making the content of the resource	string	Summation of
Size	Format.Extent		Total size of the corpus	string	total sum
Distribution Form	Type. distribution? Format. distribution?		How are the corpora distributed	open vocabulary list {CD, ftp, ... }	none
Access	Rights		The access conditions of the corpus	sub-schema	none
Pricing	Rights.price		The price of the corpus	string	none

### 3 Metadata Element Definitions

The elements for session descriptions are defined using the following attributes:

- *Element/Group Name*  
A name of the element or grouping.
- *Identifier*  
A unique identifier assigned to the element.
- *Definition*  
A statement that clearly represents the concept and essential nature of the data element.
- *Encoding*  
A statement that describes how the content of the element is encoded.
- *Comment*  
Remarks concerning the application of the data element.  
**Dublin Core equivalent:** some elements can be mapped with the Dublin Core Metadata Element Set [[DCMES](#)]. If this is possible, the Dublin Core equivalent of the IMDI element will be named here.  
**Example:** sometimes an example helps to clarify the use of the element. If this is the case, the example will be mentioned here.

#### 3.1 Catalogue

Group: Catalogue  
Identifier: Catalogue  
Definition: Groups information about a published corpus.  
Encoding: Catalogue . Name  
Catalogue . Title  
Catalogue . Id +  
Catalogue . Description +  
Catalogue . Subject Language  
Catalogue . Document Language  
Catalogue . Location  
Catalogue . Content Type  
Catalogue . Format  
Catalogue . Quality  
Catalogue . Smallest Annotation Unit  
Catalogue . Application  
Catalogue . Date  
Catalogue . Project  
Catalogue . Publisher  
Catalogue . Authors  
Catalogue . Size  
Catalogue . Distribution Form  
Catalogue . Access  
Catalogue . Pricing

Comments:

##### 3.1.1 Catalogue . Name

Element: Catalogue . Name  
Identifier: Catalogue . Name  
Definition: Name of the corpus.  
Encoding: string  
Comments:

### 3.1.2 Catalogue . Title

Element: Catalogue . Title  
Identifier: Catalogue . Title  
Definition: Title of the corpus.  
Encoding: string  
Comments:

### 3.1.3 Catalogue . Id

Element: Catalogue . Id  
Identifier: Catalogue . Id  
Definition: Unique identifier for the corpus.  
Encoding: string  
Comments: This can be an ISBN.

### 3.1.4 Catalogue . Description

Element: Catalogue . Name  
Identifier: Catalogue . Name  
Definition: Description of the corpus.  
Encoding: string  
Comments:

### 3.1.5 Catalogue . Subject Language

Element: Catalogue . Subject Language  
Identifier: Catalogue . SubjectLanguage  
Definition: Language subject of analysis.  
Encoding: See '[Language Identifier Encoding](#)' (5.1).  
Comments:

### 3.1.6 Catalogue . Document Language

Element: Catalogue . Document Language  
Identifier: Catalogue . DocumentLanguage  
Definition: Language the document is in.  
Encoding: See '[Language Identifier Encoding](#)' (5.1).  
Comments:

### 3.1.7 Catalogue . Location

Group: Catalogue . Location  
Identifier: Catalogue . Location  
Definition: Groups information about the location of where the corpus content was recorded or originated.  
Encoding: Location . Continent  
Location . Country  
Location . Region  
Comments:

#### *Catalogue . Continent*

Element: Catalogue . Continent  
Identifier: Catalogue . Continent  
Definition: The continent of where the corpus content was recorded or originated.  
Encoding: Closed controlled vocabulary { Africa, Antarctica, Asia, Australia, Europe, North America, Oceania, South America }.  
Comments:

#### *Catalogue . Country*

Element: Catalogue . Country  
Identifier: Catalogue . Country  
Definition: The country where the corpus content was recorded or originated.  
Encoding: Closed controlled vocabulary. The country is encoded with a two-letter code as described by [[ISO3166-1](#)].

Comments:

***Catalogue . Region***

Element: Catalogue . Region  
Identifier: Catalogue . Region  
Definition: The region or sub-region of where the corpus content was recorded or originated.  
Encoding: string  
Comments: This element can also be used to describe sub-regions.  
Examples: europe, the netherlands, gelderland, achterhoek.

**3.1.8 Catalogue . Content Type**

Element: Catalogue . Content Type  
Identifier: Catalogue . ContentType  
Definition: The type of the corpus.  
Encoding: Open vocabulary '[Catalogue . Content . Type](#)' (4.1).  
Comments:

**3.1.9 Catalogue . Format**

Group: Catalogue . Format  
Identifier: Catalogue . Format  
Definition: Groups information about the formats used in the corpus.  
Encoding: Format . Text  
Format . Audio  
Format . Video

Comments:

***Catalogue . Format . Text***

Element: Catalogue . Format . Text  
Identifier: Catalogue . Format . Text  
Definition: The format of the text used in the corpus.  
Encoding: string  
Comments:

***Catalogue . Format . Audio***

Element: Catalogue . Format . Audio  
Identifier: Catalogue . Format . Audio  
Definition: The format of the audio data used in the corpus.  
Encoding: string  
Comments:

***Catalogue . Format . Video***

Element: Catalogue . Format . Video  
Identifier: Catalogue . Format . Video  
Definition: The format of the video data used in the corpus.  
Encoding: string  
Comments:

**3.1.10 Catalogue . Quality**

Group: Catalogue . Quality  
Identifier: Catalogue . Quality  
Definition: Groups information about the quality of the corpus content.  
Encoding: string  
Comments:

***Catalogue . Quality . Audio***

Element: Catalogue . Quality . Audio  
Identifier: Catalogue . Quality . Audio  
Definition: The quality of the audio data in the corpus.

Encoding: Closed controlled vocabulary { 1 .. 5 }.  
Comments:

#### ***Catalogue . Quality . Video***

Element: Catalogue . Quality . Video  
Identifier: Catalogue . Quality . Video  
Definition: The quality of the video data in the corpus.  
Encoding: Closed controlled vocabulary { 1 .. 5 }.  
Comments:

#### **3.1.11 Catalogue . Smallest Annotation Unit**

Element: Catalogue . Smallest Annotation Unit  
Identifier: Catalogue . SmallestAnnotationUnit  
Definition: The smallest annotation unit used in the corpus.  
Encoding: Open vocabulary '[Catalogue . Smallest Annotation Unit](#)' (4.2).  
Comments:

#### **3.1.12 Catalogue . Application**

Element: Catalogue . Application  
Identifier: Catalogue . Application  
Definition: The application domain of the corpus.  
Encoding: Open vocabulary list '[Catalogue . Application](#)' (4.3).  
Comments:

#### **3.1.13 Catalogue . Date**

Element: Catalogue . Date  
Identifier: Catalogue . Date  
Definition: The publishing date of the corpus.  
Encoding: The date is encoded according to a profile of [[ISO8601](#)] as described in [[W3CDTF](#)] and follows the YYYY-MM-DD format.  
Comments:

#### **3.1.14 Catalogue . Project**

Element: Catalogue . Project  
Identifier: Catalogue . Project  
Definition: Name of the project for which the corpus was originally created.  
Encoding: string  
Comments:

#### **3.1.15 Catalogue . Publisher**

Element: Catalogue . Publisher  
Identifier: Catalogue . Publisher  
Definition: An entity responsible for making the resource available.  
Encoding: string  
Comments:

#### **3.1.16 Catalogue . Authors**

Element: Catalogue . Authors  
Identifier: Catalogue . Authors  
Definition: An entity primarily responsible for making the content of the resource.  
Encoding: string  
Comments:

#### **3.1.17 Catalogue . Size**

Element: Catalogue . Size  
Identifier: Catalogue . Size  
Definition: Total size of the corpus.  
Encoding: string  
Comments:

### **3.1.18 Catalogue . Distribution Form**

Element: Catalogue . Distribution Form  
Identifier: Catalogue . DistributionForm  
Definition: How are the corpora distributed.  
Encoding: string  
Comments:

### **3.1.19 Catalogue . Access**

Group: Catalogue . Access  
Identifier: Catalogue . Access  
Definition: Groups information about access rights.  
Encoding: Access (sub-schema)  
Comments:

### **3.1.20 Catalogue . Pricing**

Element: Catalogue . Pricing  
Identifier: Catalogue . Pricing  
Definition: The price of the corpus.  
Encoding: string  
Comments:

## **3.2 Sub-schemas**

### **3.2.1 Access**

See the document 'Metadata Elements for Session Descriptions' for the details.

## 4 Vocabularies

### 4.1 Catalogue . Content Type

Open vocabulary:

- Written
- Speech
- Terminology

Subtype:

- Corpus
- Monolingual Lexicon
- Multilingual Lexicon
- Telephone speech
- Desktop/microphone
- Multimodal/Multimedia
- Other speech related

For example see [\[ELRA\]](#).

### 4.2 Catalogue . Smallest Annotation Unit

Open vocabulary:

- Paragraph
- Utterance
- Word
- Phoneme
- ...

### 4.3 Catalogue . Application

Open vocabulary list. For example see the [\[LDC\]](#) set.

## 5 Encoding Formats

### 5.1 Language Identifier Encoding

The language identifier is encoded as follows:

<namespace identifier>:<language identifier>

The following namespace identifiers are allowed:

#### **ISO639-1**

Specifies the code set for language identification in the form of a two-letter code. See [\[ISO639-1\]](#).

#### **ISO639-2**

Specifies the code set for language identification in the form of a three-letter code. See [\[ISO639-2\]](#).

#### **ISO639**

Allows both [\[ISO639-1\]](#) and [\[ISO639-2\]](#) code sets for language identification.

#### **RFC1766**

Allows both two-letter [\[ISO639-1\]](#) codes and [\[ISO639-1\]](#) combined with [\[ISO3166-1\]](#) country codes. See [\[RFC1766\]](#).

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.

Examples:

ISO639-2: ger	<i>German as specified by ISO639-2</i>
RFC1766: en-US	<i>English as spoken in the US specified by RFC1766</i>
RFC1766: x-sil-dut	<i>Dutch as specified in the <a href="#">[ETHNOLOGUE]</a> list.</i>

## 6 References

- [DCMES] Dublin Core Metadata Element Set  
<http://dublincore.org/documents/dces/>
- [ETHNOLOGUE] Ethnologue language name index  
<http://www.sil.org/ethnologue/names/>
- [ELRA] European Language Resources Association, <http://www.icp.grenet.fr/ELRA/>
- [ISO639-1] Code for the representation of names of languages, International Organization for Standardization (ISO), 1988.
- [ISO639-2]  
Codes for the representation of names of languages - part 2: alpha-3 code, International Organization for Standardization (ISO), 1998.  
<http://lcweb.loc.gov/standards/iso639-2/langhome.html>
- [ISO3166-1]  
Codes for the representation of names of countries, International Organization for Standardization (ISO), 1997.  
<http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/index.html>
- [ISO8601] Data elements and interchange formats - Information interchange - Representation of dates and times, International Organization for Standardization (ISO), 2000.
- [LANGID] Language Identification and IT: Addressing problems of linguistic diversity on a global scale, Peter Constable and Gary Simons, SIL International, 2000.  
<http://www.sil.org/silewp/2000/001/>
- [LDC] Linguistic Data Consortium, <http://morph ldc.upenn.edu/>
- [OAI] Open Archives Initiative, <http://www.openarchives.org/>
- [OLAC] Open Language Archives Community, <http://www.language-archives.org/>
- [RFC1766] Tags for the identification of language  
<http://www.ietf.org/rfc/rfc1766.txt>  
specifies a two letter code taken from [ISO639-1], followed optionally by a two letter country code taken from [ISO3166-1]
- [W3CDTF] Date and Time Formats, W3C Note  
<http://www.w3.org/TR/NOTE-datetime>

## Appendix A : Catalogue Metadata Inventory

	UHLCS	ELRA	LDC
<b>Name</b>	Corpus Title	Item Name	Item Name
<b>Authors</b>			Authors
<b>Project</b>	Projects Sponsors		Project(s) +
<b>Date</b>		Creation date	
<b>Catalogue Id's</b>		ELRA Id	LDC Catalogue Id. NIST Catalogue Id. ISBN Catalogue Id.
<b>Type/Quality</b>		Type = { <i>written, speech, terminology, Tools &amp; software</i> } Subtype = { <i>Corpus, Monoling. Lex., Multiling. Lex., Telephone speech, Desktop/microphone, Mutimodal/Multimedia, Other speech related</i> }	Data type = { <i>Lexicon, Speech, Text</i> } Data source = { <i>Broadcast, conversation, microphone, mobile-radio, newswire, parallel, pronunciation, telephone, varied</i> }
<b>Format/Quality</b>			Sample Frequency Sample Format
<b>Subject Language</b>	Main-language Languages	Language	Language(s) +
<b>Documentation</b>	Description of the corpus		Online documentation Readme file
<b>Application</b>			Application={ <i>discourse analysis, information retrieval, language identification, language modelling, machine translation, message understanding, natural language processing, parsing, pronunciation modeling, speaker identification, speech recognition, speech synthesis, spoken dialogue systems, prosody, tagging, topic detection &amp; tracking</i> }
<b>Media</b>	Physical Storage format Server OS type Exploitation tools		Number of CD's (or) ftp
<b>Distributor</b>	Distributor name, address	Implicit	Implicit
<b>Pricing</b>		Member price Non-member price	Member price Non-member price Membership year (when free for members)
<b>Access/ Licensing</b>	Location Access how-to		Member license Non-member license

## Appendix B : Revision History

**Version: 2.1**

*Date: 8 June 2001; MPI ISLE Team*

First frozen element set.

**Version: 2.0**

*Date: 4 June 2001; MPI ISLE Team*

Smallest Annotation Unit was added  
Size of Corpus was added

**Version: 1.0**

*Date: 23 February 2001; MPI ISLE Team*

First version