

ISLE Metadata Initiative (IMDI)

PART 1 C

Metadata Elements for Lexicon Descriptions

Draft Proposal Version 1.1c

MPI Internal version. Next public version will be 3.0.0

IMDI Team, July 2003

INDEX

1	INTRODUCTION	3
1.1	LEXICAL METADATA	3
1.2	SCOPE OF THE PROPOSAL	3
2	LEXICON ELEMENTS OVERVIEW	4
2.1	LEXICON RESOURCE	4
2.2	LEXICON ENTRY	4
3	METADATA ELEMENT DEFINITIONS	5
3.1	LEXICON RESOURCE	5
3.2	LEXICON ENTRY	8
4	VOCABULARIES	10
4.1	LEXICON RESOURCE . TYPE	10
4.2	LEXICON RESOURCE . FORMAT	10
4.3	LEXICON ENTRY . HEADWORD TYPE	10
4.4	LEXICON ENTRY . ORTHOGRAPHY	10
4.5	LEXICON ENTRY . MORPHOLOGY	11
4.6	LEXICON ENTRY . MORPHOSYNTAX	11
4.7	LEXICON ENTRY . SYNTAX	12
4.8	LEXICON ENTRY . PHONOLOGY	12
4.9	LEXICON ENTRY . SEMANTICS	13
4.10	LEXICON ENTRY . USAGE	14
5	REFERENCES	15

1 Introduction

1.1 Lexical Metadata

.... New text

1.2 Scope of the proposal

Firstly we have to describe the scope of the metadata set discussed in this paper, since there are many types of lexical resources such as wordlists, dictionaries, glossaries, concordances, terminology databases, thesauri, and ontologies. At first instance we want to restrict ourselves to those databases which have as main entry a lexical headword and describe its characteristics. Therefore we exclude concept oriented databases such as thesauri and ontologies which relate the concept entry to other concepts in a language. It may be necessary to add other descriptive elements.

We do not want to distinguish between monolingual and multilingual lexica because some of the multilingual lexica can be broken down into monolingual lexica and a special list containing the SynSets as in Wordnet. Other lexica may have multilingual entries as part of their structure. In this case the content description will describe this appropriately.

2 Lexicon Elements Overview

2.1 Lexicon Resource

Lexicon Resource	
ResourceLink (c)	
MediaResourceLink (c)	
Date (c)	
Type (ov)	
Format (ov)	
Character Encoding (string)	
SchemaRef (c)	
Size (c)	
Number of Head Entries (c)	
Number of Sub Entries (c)	
Lexicon Entries (c)	
Meta Languages (group)	
	Language (sub) *
	Description (sub)
Access (sub)	
Description (sub)	
Keys (sub)	

2.2 Lexicon Entry

Lexicon Entry	
HeadwordType (ov)	
Orthography (ov)	
Morphology (ov)	
MorphoSyntax (ov)	
Syntax (ov)	
Phonology (ov)	
Semantics (ov)	
Etymology (string)	
Usage (ov)	
Frequency (string)	

Legend	
*	indicates a list of zero or more elements
+	indicates a list of one or more elements
String	sequence of alphanumeric symbols including spaces and punctuation
Sub	sub-schema
Group	grouping of elements
C	the element is constrained by a certain encoding scheme
Ccv	closed controlled vocabulary - the content of the element must be selected from a closed set of values.
Ov	open vocabulary - the content of the element can be selected from a predefined set of suggested values or can be user defined. An ov can later be changed into a ccv provided by some repository
Ovl	open vocabulary list - a list of values for the content of the element can be selected from a predefined set of suggested values or can be user defined. An ov can later be changed into a ccv provided by some repository

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.¹

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 Lexicon Resource

Group: Lexicon Resource
Identifier: LexiconResource
Definition: **Definition needed here!**
Encoding: LexiconResource . ResourceLink
LexiconResource . MediaResourceLink
LexiconResource . Date
LexiconResource . Type
LexiconResource . Format
LexiconResource . Character Encoding
LexiconResource . SchemaRef
LexiconResource . Size
LexiconResource . Number of Head Entries
LexiconResource . Number of Sub Entries
LexiconResource . Lexicon Entries
LexiconResource . Meta Languages
LexiconResource . Access
LexiconResource . Description
LexiconResource . Keys

Comments:

3.1.1 Lexicon Resource . ResourceLink

Group: Lexicon Resource . ResourceLink
Identifier: LexiconResource . ResourceLink
Definition: Link to the lexical resource.
Encoding: The link is encoded as a Uniform Resource Locator as described by [[RFC1738](#)]

Comments:

¹ The mapping of IMDI elements to DC elements is done here in a simplified way. While IMDI elements are embedded in a structure, DC only describes a flat list of elements. The consequences of structure are ignored here to keep the mapping simple. More careful statements about IMDI - DC mapping will be made in a follow-up document.

3.1.2 Lexicon Resource . Media Resource Link

Group: Lexicon Resource . Media Resource Link
Identifier: LexiconResource . MediaResourceLink
Definition: A link to the media resource connected to the lexical resource.
Encoding: The link is encoded as a Uniform Resource Locator as described by [\[RFC1738\]](#)
Comments: **Incomplete definition**

3.1.3 Lexicon Resource . Date

Group: Lexicon Resource . Date
Identifier: LexiconResource . Date
Definition: Date when the lexical resource was created
Encoding:
Comments:

3.1.4 Lexicon Resource . Type

Group: Lexicon Resource . Type
Identifier: LexiconResource . Type
Definition: The type of the lexical resource.
Encoding: Open vocabulary '[Lexicon Resource . Type](#)'.
Comments:

3.1.5 Lexicon Resource . Format

Group: Lexicon Resource . Format
Identifier: LexiconResource . Format
Definition: The format of the lexical resource.
Encoding: Open vocabulary '[Lexicon Resource . Format](#)'.
Comments:

3.1.6 Lexicon Resource . Character Encoding

Group: Lexicon Resource . Character Encoding
Identifier: LexiconResource . CharacterEncoding
Definition: Name of the character encoding used in the lexical resource.
Encoding: String
Comments:

3.1.7 Lexicon Resource . SchemaRef

Group: Lexicon Resource . SchemaRef
Identifier: LexiconResource . SchemaRef
Definition: A reference to a structure schema for the lexical resource.
Encoding: The link is encoded as a Uniform Resource Locator as described by [\[RFC1738\]](#)
Comments:

3.1.8 Lexicon Resource . Size

Group: Lexicon Resource . Size
Identifier: LexiconResource . Size
Definition: The size of the lexical resource.
Encoding: Number
Comments: **Machine processed?**

3.1.9 Lexicon Resource . Number of Head Entries

Group: Lexicon Resource . Number of Head Entries
Identifier: LexiconResource . NoHeadEntries
Definition: The number of head entries of the lexical resource.
Encoding: Number
Comments: **Incomplete definition**

3.1.10 Lexicon Resource . Number of Sub Entries

Group: Lexicon Resource . Number of Sub Entries
Identifier: LexiconResource . NoSubEntries
Definition: The number of sub entries of the lexical resource.
Encoding: Number
Comments: **Incomplete definition**

3.1.11 Lexicon Resource . Lexicon Entries

Group: Lexicon Resource . Lexicon Entries
Identifier: LexiconResource . Lexicon Entries
Definition: List of lexical entries from the lexical resource.
Encoding: Lexicon Entries . Lexicon Entry *
Comments:

3.1.12 Lexicon Resource . Meta Languages

Group: Lexicon Resource . Meta Languages
Identifier: LexiconResource . MetaLanguages
Definition: A block to describe the languages that are used to define terms, to describe meaning.
Encoding: Meta Languages . Language *
Meta Languages . Description
Comments: **Incomplete definition**

Meta Languages . Language

Group: Meta Languages . Language
Identifier: MetaLanguages . Language
Definition: The language used to define terms, to describe meaning.
Encoding: Language (sub-schema)
Comments: **Incomplete definition**

Meta Languages . Description

Group: Meta Languages . Description
Identifier: MetaLanguages . Description
Definition: -
Encoding: Description (sub-schema)
Comments: **Incomplete definition**

3.1.15 Lexicon Resource . Access

Group: Lexicon Resource . Access
Identifier: LexiconResource . Access
Definition: The access rights of the lexical resource.
Encoding: Access (sub-schema)
Comments:

3.1.16 Lexicon Resource . Description

Group: Lexicon Resource . Description
Identifier: LexiconResource . Description
Definition: A description of the lexical resource.
Encoding: Description (sub-schema)
Comments:

3.1.17 Lexicon Resource . Keys

Group: Lexicon Resource . Keys
Identifier: LexiconResource . Keys
Definition: Name-value pairs to describe domain specific information about the lexical resource.
Encoding: Keys (sub-schema)
Comments:

3.2 Lexicon Entry

The lexicon entry elements describe the linguistic content covered by each lexicon.

The following main entry categories have been distinguished as a proposal for implementation. Each main category contains zero or more names of object classes that represent subclasses of the linguistic descriptive level captured by the category. Each occurrence indicates whether the corresponding linguistic information is present in the lexicon. Some of these object classes can have subclasses of their own which are not shown and discussed here. The idea is that at a later moment not only the names of these object classes will be available, but that the user can receive more detailed information. The list of object classes is not meant to be exhaustive and can be extended if necessary.

In order to accommodate linguistic annotation in a maximally polytheoretic and flexible fashion, it is possible to duplicate existing subclasses as descendants of other main categories if the need arises. The categories do not make any statement about the details of the encoding and whether subdivisions are used. The purpose is that the person searching for e.g. morphological segmentation data will find a hit in the meta-description of some lexicon, and subsequently has to take a more detailed look into the resource itself to find out about the format and granularity of the available segmentation data.

Group: Lexicon Entry
Identifier: LexiconEntry
Definition: The linguistic content of the lexicon.
Encoding: Lexicon Entry . HeadwordType
Lexicon Entry . Orthography
Lexicon Entry . Morphology
Lexicon Entry . MorphoSyntax
Lexicon Entry . Syntax
Lexicon Entry . Phonology
Lexicon Entry . Semantics
Lexicon Entry . Etymology
Lexicon Entry . Usage
Lexicon Entry . Frequency
Comments: The attributes the lexicon contains.

3.2.1 Lexicon Entry . Headword Type

Group: Lexicon Entry . Headword Type
Identifier: LexiconEntry . HeadwordType
Definition: The linguistic nature of the entry in the lexicon.
Encoding: Open vocabulary '[Headword Type](#)'.
Comments:

3.2.2 Lexicon Entry . Orthography

Group: Lexicon Entry . Orthography
Identifier: LexiconEntry . Orthography
Definition: Orthography used in the lexical resource.
Encoding: Open vocabulary '[Orthography](#)'.
Comments:

3.2.3 Lexicon Entry . Morphology

Group: Lexicon Entry . Morphology
Identifier: LexiconEntry . Morphology
Definition: Morphology used in the lexical resource.
Encoding: Open vocabulary '[Morphology](#)'.
Comments:

3.2.4 Lexicon Entry . MorphoSyntax

Group: Lexicon Entry . MorphoSyntax
Identifier: LexiconEntry . MorphoSyntax
Definition: Morphosyntax used in the lexical resource.
Encoding: Open vocabulary '[MorphoSyntax](#)'.
Comments:

3.2.5 Lexicon Entry . Syntax

Group: Lexicon Entry . Syntax
Identifier: LexiconEntry . Syntax
Definition: Syntax used in the lexical resource.
Encoding: Open vocabulary '[Syntax](#)'.
Comments:

3.2.6 Lexicon Entry . Phonology

Group: Lexicon Entry . Phonology
Identifier: LexiconEntry . Phonology
Definition: Phonology used in the lexical resource.
Encoding: Open vocabulary '[Phonology](#)'.
Comments:

3.2.7 Lexicon Entry . Semantics

Group: Lexicon Entry . Semantics
Identifier: LexiconEntry . Semantics
Definition: Semantics used in the lexical resource.
Encoding: Open vocabulary '[Semantics](#)'.
Comments:

3.2.8 Lexicon Entry . Etymology

Group: Lexicon Entry . Etymology
Identifier: LexiconEntry . Etymology
Definition: Information about the historical context of a lexical entry or wordform.
Encoding: String
Comments: E.g. morphological, phonological, syntactic, semantic.

3.2.9 Lexicon Entry . Usage

Group: Lexicon Entry . Usage
Identifier: LexiconEntry . Usage
Definition: Pragmatic/sociolinguistic information.
Encoding: Open vocabulary '[Usage](#)'.
Comments: **Definition incomplete**

3.2.10 Lexicon Entry . Frequency

Group: Lexicon Entry . Frequency
Identifier: LexiconEntry . Frequency
Definition: Corpus-derived frequency of occurrence.
Encoding: String
Comments: **Definition incomplete**

4 Vocabularies

4.1 Lexicon Resource . Type

Possible values are:

- Dictionary
- Wordlist
- Glossary
- Concordance
- Terminology

Exact definitions to be provided

4.2 Lexicon Resource . Format

...

4.3 Lexicon Entry . Headword Type

Possible values are:

- Sentence
- Phrase
- Wordform
- Lemma
- Abstract Lemma
- Stem
- Affix

Exact definitions to be provided

Value: Sentence

Definition: -

Value: Phrase

Definition: -

Value: Wordform

Definition: -

Value: Lemma

Definition: The entry conforms to the unmarked wordform.

Examples: Infinitive for verbs.

Value: Abstract Lemma

Definition: The entry does not conform to any wordform of the group subsumed by the lemma.

Value: Stem

Definition: -

Value: Affix

Definition: -

4.4 Lexicon Entry . Orthography

Possible values are:

- Hyphenated Spelling
- Syllabified Spelling

- Spelling Variants
- Citations

Exact definitions to be provided

Value: Hyphenated Spelling
 Definition: -

Value: Syllabified Spelling
 Definition: -

Value: Spelling Variants
 Definition: Orthographic variations with or without preferred spelling information.

Value: Citations
 Definition: -

4.5 Lexicon Entry . Morphology

Possible values are:

- Stem
- Stem Allomorphy
- Segmentation
- Production rules
- Typology

Definitions to be verified

Value: Stem
 Definition: Deep or surface stem.

Value: Stem Allomorphy
 Definition: Variations at stem level.

Value: Segmentation
 Definition: Analysis into morphological constituents such as affixes.

Value: Production rules
 Definition: Governing the production of surface forms on the basis of stems.

Value: Typology
 Definition: Any classification of entries or morphological entities.

4.6 Lexicon Entry . MorphoSyntax

possible values are:

- Part of Speech
- Inflection
- Countability
- Gradability
- Gender
- Typology

Definitions to be verified

Value: Part of Speech
 Definition: Syntactic class of the entry.

Value: Inflection
 Definition: Any inflectional or conjugational information.

Value: Countability
Definition: Pluralization properties.

Value: Gradability
Definition: -
Examples: Adjectival comparative/superlative constructions.

Value: Gender
Definition: -
Examples: Neuter

Value: Typology
Definition: Any classification of entries.

4.7 Lexicon Entry . Syntax

possible values are:

- Complementation
- Alternation
- Modification
- Shallow Parsing
- Deep Parsing
- Functional Parsing
- Collocations
- Typology

Definitions to be verified

Value: Complementation
Definition: Syntactic subcategorization.

Value: Alternation
Definition: Alternative complementation patterns.

Value: Modification
Definition: -
Examples: Adjectival modification patterns.

Value: Shallow Parsing
Definition: Segmentation into chunks.

Value: Deep Parsing
Definition: Finer grained analysis below chunk level.

Value: Functional Parsing
Definition: Syntactic functions such as subject.

Value: Collocations
Definition: Significant juxtaposed entries/wordforms.

Value: Typology
Definition: Any classification.
Examples: Prepositional/phrasal verb.

4.8 Lexicon Entry . Phonology

possible values are:

- Transcription
- IPA Transcription

- CV pattern
- Constituent Structure
- Intonation

Definitions to be verified

Value: Transcription
 Definition: Any type of phonetic/phonological transcription.

Value: IPA Transcription
 Definition: Transcription in International Phonetic Alphabet.

Value: CV pattern
 Definition: Transcription in terms of consonant-vocal combinations.

Value: Constituent Structure
 Definition: Segmentation in to phonetic constituents.

Value: Intonation
 Definition: -
 Examples: Stress marking, constituent length etc.

4.9 Lexicon Entry . Semantics

possible values are:

- Sense distinction
- Ontological classification
- Gloss
- Definition
- Connotation
- Idiom
- Componential Features
- Cross-references
- Semantic relations
- Preference

Definitions to be verified

Value: Sense distinction
 Definition: Polysemy and/or homonymy.

Value: Ontological classification
 Definition: Related concepts and conceptual relations.

Value: Gloss
 Definition: Informal description of the sense in natural language.

Value: Definition
 Definition: formal description of the sense.
 Examples: A 1st order logic formula.

Value: Connotation
 Definition: Non-denotational information such as pejorative.

Value: Idiom
 Definition: Idiosyncratic use.

Value: Componential Features.
 Definition: Formula or list containing a finite set of meaning attributes.

Value: Cross-references
Definition: Links to other entries/wordforms.

Value: Semantic relations
Definition: Relations between entries or associated concepts.

Value: Preference
Definition: Characterization of the arguments in the semantic predicate.

4.10 Lexicon Entry . Usage

possible values are:

- Region
- Style

Exact definitions to be provided

Value: Region
Definition: -
Examples: Dialect.

Value: Style
Definition: -
Examples: Slang

5 References

- Bell, J., Bird, St. A Preliminary Study if the Structure of Lexicon Entries.
<http://www ldc.upenn.edu/exploration/exp12000/program.html>
- [CELEX]
<http://www.mpi.nl/world/celex/>
- [DCMES] Dublin Core Metadata Element Set
<http://dublincore.org/documents/dces/>
- [DOBES] Documentation of Endangered Languages
<http://www.mpi.nl/DOBES/>
- Gibbon, D. (2001) *Notes on Lexicon Metadata*. ISLE CLWG Meeting, Pisa
- Ide, N., Romary, L. (2001) A Flexible Framework for Representing Computational Lexicons. ISLE CLWG Meeting, Pisa
- [IMDI] ISLE Metadata Initiative
<http://www.mpi.nl/IMDI/>
- Manning, C. *Kirrkirr*. Lexical Tool
<http://www-nlp.stanford.edu/kirrkirr/>
- Peters, W. (2000) *Metadata for Lexicons*. MPI Lexicon Workshop, Nijmegen
- Peters, W. (2001) *Metadata for Lexical Resources*. ISLE CLWG Meeting, Pisa
- [RFC1738] Uniform Resource Locators
<http://www.w3.org/Addressing/rfc1738.txt>
- [TEI] Text Encoding Initiative
<http://www.tei-c.org/>
- [W3CDTF] Date and Time Formats, W3C Note
<http://www.w3.org/TR/NOTE-datetime>
- Wittenburg, P. (2001) *Lexical Structures*. DOBES internal document. MPI Nijmegen