ISLE Metadata Initiative (IMDI)

PART 1 C

Metadata Elements for Lexicon Descriptions

Draft Proposal Version 1.0

December, 2001

IMDI[1] Technical Report

Peter Wittenburg
Dafydd Gibbon
Wim Peters

[1] For information about the ISLE Metadata Initiative, please, look at the following web-site: www.mpi.nl/ISLE

INDEX

1 Introduction

Within ISLE a work package was defined which has to deliver, among other things metadata standards for multimedia/multimodal language resources. The ISLE MetaData Initiative (IMDI) [1] was formed and many outstanding scientists from the discipline have joined the boards. The IMDI Steering Board decided at its Athens meeting [see 1] that the initiative should restrict itself to developing metadata sets for multimedia/multimodal corpora and lexica. With respect to corpora IMDI established a metadata set and developed tools [see 1]. With respect to lexica it was stated that a first draft proposal should be ready at the end of 2001. With this paper the authors wish to present their ideas about a lexicon metadata set to the field experts.

1.1 Lexical Metadata

Before discussing the task in more detail it is necessary to explain the term "metadata" in the context of lexica. There are often statements such as "all data is metadata" or "lexical data is metadata". Dependent on the perspective both visions are correct, of course. However, when we speak about lexical metadata in this paper we mean data which describes lexical resources which can be easily retrieved suitable ones with the help of simple queries. Even this view does not fully clarify what lexical metadata is, since experts probably have much more detailed questions for finding resources compared to those who are simply looking for an English translation of a German word.

With respect to metadata there are two major differences between corpora and lexica resources. (1) Lexical data are abstractions from corpora data or language use[2]. This means that finding a suitable resource is far less problematic. While there will be hundreds to thousands of annotated recordings for a certain language there will be comparatively few lexical resources. (2) One could argue that given the existence of a schema description for a lexical resource this schema itself describes its linguistic content at a high level of detail. So the tags used in the schema describe the content of the lexicon. If well-agreed tag labels were used this might already be sufficient. However, there are many lexica which don't have a schema description associated with them and many elements seen as being necessary in metadata descriptions (see chapter 3) are not included.

Since we are faced mostly with lexica for which there is no schema definition, since the schema definitions in many cases will offer too much detail, since the lexicon structures are very different from each other as a recent investigation showed [2] and since the tag labels used by different creators are still very different despite all standardization efforts, the authors stick to the idea to create a separate standard for lexical metadata descriptions. Of course, this standard has to correspond closely to what is currently discussed in the various initiatives trying to define generic standards for lexica.

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

[2] Due to this abstraction process some people speak of lexical data as metadata in a general sense.

a lexical headword and describe its characteristics. Therefore we will 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.

1.3 Lexicon Initiatives

Much work has already been carried out on standardizing lexicons especially to facilitate language engineering applications. While TEI [3] does not make detailed proposals for lexical tag sets various projects often executed under the EAGLES/ISLE [4] umbrella worked out concrete lexical structures. GENELEX [5] can be seen as an early attempt to describe a generic lexicon structure with an exhaustive DTD including many attributed. SIMPLE [6] was an attempt to encode multilingual lexica in a uniform way with the result of a 12 smaller example lexica. MULTILEX [7] was another project focusing on the implementation of 15 concrete lexica applying a structure derived from the GENELEX model. The MILE (Multilingual Computational Lexicon) project [8] recently started within ISLE has the task of standardizing multilingual lexica.

Other relevant work was undertaken by the OLIF2 consortium resulting in the OLIF2 proposal (Open Lexicon Interchange Format) [9]. OLIF2 defines a large number of lexical features, but does not make statements about their structural embedding. Each OLIF2 entry is a monolingual entry containing various feature/value pairs, cross-references between entries in the same language lexicon, and transfers defining bilingual transfer relations. The OLIF2 proposal for features describes four main categories: administrative, morphological, syntactic, semantic. The features are similar to those found in other more generic lexicon proposals.

Much work has been done in the area of terminology databases. The MARTIF (Machine Reachable Terminology Interchange Format) [10] work describes a format to facilitate the interchange of terminological data among terminology management systems. This work resulted in the ISO 12200 specifications. Complementary to that ISO 12620 [11] specifies how “Data Categories” which are the basic elements to describe lexical content have to be defined. Lexical terms can be defined by designating a concept entry and being associated with a language. Term related information specifies the type of the terms added. This is done by assigning attributes to the entries such as POS, etymology etc which are labels also found in dictionaries serving purposes other than terminology. Description information sets the terms into relation with domains and points to positions in concept hierarchies. Administrative information can also be added to each term. Among other things it includes typical metadata such as creator name, creation date etc.

Another important initiative is CLS (Concept Oriented with Links and Shared references) [12]. While ISO 12620 only defines the concept an entry belongs to, CLS also defines the structures between data items within entries. In doing so CLS is a framework for defining structure and content of terminology databases. The SALT project (Standards-based Access to Lexicon and Terminologies [13]) was recently initiated mainly driven by the needs from language engineering. SALT suggests the

XLT (XML representations of Lexicons and Terminologies) family of formats for representing, manipulating, and sharing terminological data. A core DTD/Schema defining the generic structure can be combined with a particular Data Constraint Specification which results in a member of the XLT family of formats. The data constraint specification determines which data categories will be used in the particular application and which values these can have. The core structure of SALT is based on the MARTIF proposal.

As shown, a number of these projects introduce categories which can also be seen as typical metadata. OLIF2 presents the most detailed list. However, none of these projects addressed the problem of finding lexical resources on the Web by querying on structured metadata descriptions in particular. Since it will be important to automatically extract metadata categories from the term descriptions, for example in terminology databases, it will be necessary to closely synchronize with the projects mentioned.

1.4 Metadata, XML and RDF

Most of the metadata sets published so far are simply lists of descriptor elements which can be used once or several times to characterize a resource. An example is Dublin Core which defines 15 elements, some qualifiers to refine the elements and some constraints. An example of a structure supporting metadata standard is IMDI. The reasons for supporting structure were to allow the definition of dependencies and in doing so to make queries possible which are of interest to the users.

The following simple example shows that in the structured case it is possible to answer queries such as: “Give me all resources where the age of participants is ‘8’ and where they speak ‘German’”. This is not possible in the unstructured case.

Structured description such as with IMDI	Set-like description such as with DC
participant	contributor=peter
participant.name=peter	contributor=jan
participant.age=6	age=6
participant.language=english	age=8
participant	language=english
participant.name=jan	language=german
participant.age=8	
participant.language=german	

Of course the creators of the DC set knew this problem. RDF (Resource Description Framework) [14] was designed to combine metadata elements from different sets which are stored in some repository such as DC in http://purl.org/DC/ and to relate these elements to each other. RDF syntax allows us also to introduce structure to cope with problems as indicated in the example above.

The need to be able to combine elements from different sets goes back to the Warwick framework [15]. It was recognized that there will be many metadata sets dedicated to domains and even groups within domains. The idea was and remains that there should be mechanisms to combine semantic definitions (elements from the vocabularies) from several metadata descriptions to create a new more complex metadata standard. RDF was amongst other reasons developed to allow this combination. In RDF one first has to specify the name spaces which will be used by referring to URIs where the corresponding repositories can be found. One can then reuse elements from the different repositories for one’s own purposes.

RDF makes use of XML as underlying syntax and was designed before the XML Schema idea arose. There is so much overlap on the structural level that RDF can be seen especially as a framework to specify semantic relationships between elements.

1.5 Approach

In this document we will follow the approach used within the IMDI project. (1) We will develop a metadata set as required by the community and which can be implemented soon. The definition will be available via a public URI. (2) We will describe mappings to other metadata sets such as DC to enable OAI type of metadata harvesting (see www.mpi.nl/ISLE for an IMDI to DC mapping). (3) The vocabulary will be entered into the SALT terminology database.

The described approach is necessary as we first have to understand what the user community requires. Starting from DC as some other initiatives have done does not lead to the necessary understanding of the domain. This approach was also chosen by the MPEG7 initiative.

In the case of exact mappings of the terms used for example in IMDI and in DC it will be possible later to create RDF-based schema definitions of the IMDI sets. This would open another way to achieve interoperability between applications which may evolve during the coming decade. Currently, OAI type of harvesting appears sufficient to satisfy the needs of a very broad community.

2 Metadata Suggestions

During the past year a number of presentations were given which make suggestions about the metadata elements which should be used to describe lexica.

2.1 Peters' Proposal

Wim Peters presented two papers one in September 2000 [16] and another in February 2001 [17]. In the first paper he looks at lexicons having lemma based entries which are associated with linguistic descriptions from 5 main areas: orthography, syntax, morphology, semantics, phonology. Another important area is usage information. Of course, these main areas can be divided into sub-categories which will not be listed here. The way sub-categorial information is represented is different for the various lexica. He also discusses that lexica can be in different formats such as XML-structures, typed-feature structures, relational structures, or project specific idiosyncratic structures. Finally, he refers to the work started within the GENELEX and ISLE projects which try to homogenize the linguistic information in lexical resources. A checklist was created for evaluation purposes which can also serve as a source for defining relevant metadata characteristics. The following slots where defined in a tree-like structure:

1 Headword; lemma; entry
2 phonetic transcription
3 alternative spelling
4 inflection, conjugation
5 cross-reference
6 morphosyntax
6.1 POS
6.2 inflectional class
6.3 derivation
6.4 gender
6.5 number
6.6 mass vs count
6.7 gradation
7 subdivision counter
8 entry subdivision
9 sense indicator
10 linguistic label
11.1 subcategorization frame
11.2 obligatory of elements
11.3 auxiliary
11.4 light or support construction
11.5 periphrastic constructions
11.6 phrasal verbs
11.7 collocator
11.8 alternations
12.1 semantic type
12.2 argument structure
12.3 semantic relations
12.4 regular polysemy
12.5 domain
12.6 decomposition
13 translation
14 gloss; definition
15 near-equivalent
16 example phrase (straightforward)
17 example phrase (problematic)
18 multiword unit
19 subheadword
20 usage note
21 frequency

In his second paper a step towards formulating lexical metadata was made. He distinguishes external (creator, associated tools, ...) and internal information (linguistic content). With respect to external information Peters suggests including the following elements: ID, Date of creation, Title, Creator, Publisher, Contributor, Project, Contact, Format[3], URL, Type[4], Access, Source, Relation, References. With respect to internal information, Peters mentions the following descriptors: Content-

[3] oracle/MS access/shoebox/...
[4] collection, dataset, software/ ...

Description, Content-Language, Content-Size, Region, Range, Modality[5], Data Format[6].

Further, Peters describes the differences between top-down and bottom-up developed metadata sets for lexica. The advantages of the top-down design are a high-level, underspecified and uniform set of metadata and a high degree of theory independence. The bottom-up approach starts with concrete examples and tries to establish commonalities. If the sample is large enough both approaches come very close. A combination of both approaches is advocated and the results of the MILE project are seen as relevant.

2.2 Gibbon's Proposal

Dafydd Gibbon also makes suggestions about lexical metadata in his MILE presentation [18]. He describes a layered approach to the problem of defining a metadata set for lexica. He distinguishes two levels of lexical objects: (1) The Lexicon Object covering general information about the lexicon as a whole. (2) Information about the Lexical Entries which describe the content. With respect to the Lexicon Object three categories are distinguished: (1) bibliographical data such as creator, publisher, title, date etc; (2) Medium and format aspects; (3) macrostructure type of information such as languages involved, lexicon type (taxonomy), etc.

The description of a Lexical Entry comprises a type description and a microstructure description. The latter gives information about the underlying structure of the entries and the enclosed data category groups. These are the classical type of high level linguistic information categories each having their special list of data categories defined by some user groups or standardization bodies. Gibbon further elaborates by giving an impression of how the structure of the data categories could be described.

Lexicon Object
- bibliographical data (creator, publisher, title, date, ...)
- Medium and format description
- macrostructure type (languages involved, lexicon type/taxonomy, ...)

Lexical Entry
- type (encyclopaedic, multiword unit, ...)
- microstructure data model
  - structure (flat, tree,...)
  - data category groups
    - data categories - surface
    - data categories - structure
    - data categories - meaning
    - data categories - examples
    - data categories - housekeeping

2.3 Ide&Romary Paper

A paper presented by Nancy Ide and Laurent Romary [19] does not explicitly mention the metadata issue, but due to its focus on a flexible framework for representing computational lexica it is also relevant to the metadata topic. Basically they argue for the definition of common data categories (lexical objects), their

[5] linguistic mode such as orthography, sound, gesture, ...
[6] relational, feature structure, relational, list, ...

representation with the help of RDF, and their embedding in an abstract XML representation. The description of a data category is metadata which is correct in the broader sense of metadata. Giving such definitions it would be possible to automatically extract those descriptors which are called metadata elements in this paper.

2.4 Overview about Lexicon Structures

Recently, Wittenburg presented an analysis of the structures of real lexica [2]. These were presented by the teams in the DOBES project [20], an analysis of structures of written lexica by Bell and Bird [21], and an analysis of the structures of well-known lexica such as GENELEX [5], CELEX [22], and KirrKirr [23]. It was shown that the structures used were very different and that the linguistic sub-domain is not the most relevant criterion. In field linguistics the heterogeneity in structural complexity is naturally the largest. The more complex structures in field linguistics are comparable to what is known, for example, from CELEX. The structure of written lexica is fairly simple, while the GENELEX model had to be exhaustive.

In field linguistics it can be seen that the language under investigation influences the structural choices and the complexity. In highly inflected languages such as Wichita most utterances are lexical entries bearing a rich internal structure. Documenting these languages requires a rich lexicon.

Another observation was that the grouping of data categories (the primitive elements found in lexica) is also very heterogeneous. This leads to the assumption that differences in linguistic theory and languages influence the structural choices.

3 Design for Lexical Metadata

3.1 Introductory Remarks

Already at this moment it seems to be wise to embed lexical metadata descriptions in an RDF schema. This allows us to make use of the various definitions of lexical data categories. As we have seen there are a number of projects which have started to define them. Lexical metadata descriptions have to make use of them as far as possible.

From the overview about lexical structures we can conclude that lexicon metadata can not go into too deep detail when describing the lexical entries. Lexical metadata cannot intend to represent the lexical structure, since there are so many differences and linguists in general don't agree about the details. In this proposal we have chosen an approach to mention the main categories of lexical data and to allow people to add values to these categories which represent the common names for lexical sub-categories. However, the values associated with these main categories have to be selected from an controlled vocabulary and added as a flat list. Only such an approach will promise to not create over-specifications which will prevent successful searching. Due to the differences in linguistic theory which partly represents the different natures of the languages under study sub-categories can be listed under more than one main category.

Example with two main categories:

Orthography:	Spelling, Syllabification, Hyphenation
Morphosyntax:	POS, Inflection, Gender

The semantics of this typical description can be read in the following way: The lexicon includes information about the two main categories orthography and morphosyntax. Orthography incorporates information about spelling, syllabification and hyphenation. Morphosyntax incorporates information about POS, Inflections and Gender. The entries do not make statement about the details of the encoding and whether substructures are used etc. The purpose is that the person searching for inflectional data will find a hit and then has to look in detail into the resource itself to understand how it is done.

As in the case of corpora we have to include flexibility in the metadata set such that researchers can add important fields which are not yet in the proposed element set.

3.2 Lexicon Object Elements

We would like to follow Gibbon's classification and first discuss metadata elements on the level of the lexicon as a whole. Here we let us guide by the proposals from Peters, Gibbon, the current IMDI set for describing multimedia corpora, and the suggestions made in the various lexicon and terminology projects discussed above[7].

Lexicon Object	
Name	A short name which identifies the Lexicon
Title	A more elaborated title of the Lexicon
Date+	Date of the creation and major modifications

[7] It has to be noted here that it is not mandatory to fill in all fields.

Version	Version indication
LexiconType	Type following some taxonomy[8] (controlled vocabulary)
Creator+	The responsible persons who created the resource
Name	Name of creators
Contact	Contact address which can be the creators address or a substitution address
Description	A suitable description associated with the set of creators
Project	A block to describe the project
Name	Short name of the project
ID	Unique project identifier
Contact	contact address sub schema
Description+	some space for descriptions to be associated with the project
Object Languages	A block to describe the languages included in the lexicon
Description	some space for a prose description
MultilingualityType	languages can occur in different flavors in lexica, they can occur as multilingual entries in ML lexica, but they also can occur as translations of for example sense descriptions; this difference can be indicated with the help of a controlled vocabulary
Language+	a list of languages included, each language be described in a substructure
Meta Languages	A block to describe the languages which are used to define terms, to describe meaning and similar
Description	some space for a prose description
Language+	a list of languages included, each language be described in a substructure
Lexical Entry	a block which describes the linguistic content, i.e. the attributes the lexicon contains
EntryElements	a possibility to add feature/value pairs describing the linguistic categories used in the lexicon; the feature is taken from a controlled vocabulary; per feature a flat list of descriptors can be specified also taken from a controlled vocabulary; for details see below
Format	a rough indication of the format the lexicon is in such as relational table, structured plain text, some XML format, html format, …
AccessTool	many lexica are only interpretable via concrete access tools such as Shoebox, ORACLE, FoxPro, Access, Web-Browser, …
Media	this entry tells whether the lexicon includes audio or video samples or graphics
Schema	name of or ref to the documented structure which could be a DTD, Schema or similar
Character Encoding+	this list should give an impression of the type of fonts needed to render all data included such as UTF-8, ISO-latin
Size	the size of the lexicon in bytes
No Lexical Entries	the number of lexical entries the lexicon includes
Access	sub-schema where access info is given
Keys	a possibility to add feature/value pairs to define new keywords
Source	this entry can't be anything else than a "special description" which describes which sources were used to build the lexicon; this info is relevant, but I don't see any way to make it a formal description
Description+	these are general descriptions which can be associated with the lexicon
References	block to cover references to publications etc
Description+	

Sub-Schemas

[8] This metadata set at first instance was only made for a selection of lexical types such as dictionary | wordlist | glossary | concordance | terminologyDB

Access

ResourceLink	URL pointing to the resource if it is directly accessible
Availability	codification of terms of access (has to be worked out
Description	prose description associated with access
Date	date of statements about access
Owner	defines the owner of the lexicon
Publisher	defines the publisher of the lexicon
Contact	specifies a sub-schema describing whom to contact

Contact

Name	name of the contact person
Address	address info
Email	email address
Organization	name of an institution

Description

Text	this is the prose text of the description
Language ID	this describes the language the text is in
Info Link	each description can be associated with further information such as web-pages etc

Language

Language ID	formal language specifier from ISO or SIL lists
Name	general name of the language
Description	a description of the language can be associated

3.3 Lexicon Entry Elements

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

The following 11 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.

Modality indicates which mode of communication is captured in the lexicon. Possible values are:

- Spoken
- Written
- Sign

Headword type indication of the linguistic nature of the entry in the lexicon. Possible values are:

Sentence	
Phrase	
Wordform	
Lemma	entry conforming to the unmarked wordform (e.g. infinitive for verbs).
Abstract Lemma	entry not conforming to any wordform of the group subsumed by the lemma.

Stem
Affix

Orthography possible values are:
Hyphenated Spelling
Syllabified Spelling
Spelling Variants orthographic variations with or without preferred spelling information
Citations

Morphology possible values are:

Stem	deep or surface stem
Stem Allomorphy	variations at stem level
Segmentation	analysis into morphological constituents such as affixes
Production rules	governing the production of surfaec forms on the basis of stems
Typology	any classification of entries or morphological entities

Morphosyntax possible values are:

Part of Speech	syntactic class of the entry.
Inflection	any inflectional or conjugational information
Countability	pluralization properties
Gradability	e.g. adjectival comparative/superlative constructions
Gender	e.g. neuter
Typology	any classification of entries

Syntax possible values are:

Complementation	Syntactic subcategorization
Alternation	alternative complementation patterns
Modification	e.g. adjectival modification patterns
Shallow Parsing	segmentation into chunks
Deep Parsing	finer grained analysis below chunk level
Functional Parsing	syntactic functions such as subject
Collocations	significant juxtaposed entries/wordforms
Typology	any classification, e.g. prepositional/phrasal verb

Phonology possible values are:

Transcription	any type of phonetic/phonological transcription
IPA Transcription	transcription in International Phonetic Alphabet
CV pattern	transcription in terms of consonant-vocal combinations
Constituent Structure	segmentation in to phonetic constituents
Intonation	stress marking, constituent length etc.

Semantics possible values are:

Sense distinction	polysemy and/or homonymy
Ontological classification	related concepts and conceptual relations
Gloss	informal description of the sense in natural language
Definition	formal description of the sense e.g. as a 1st order logic formula
Connotation	non-denotational information such as pejorative
Idiom	idiosyncratic use
Componential Features	formula or list containing a finite set of meaning attributes
Cross-references	links to other entries/wordforms
Semantic relations	relations between entries or associated concepts
Preference	characterization of the arguments in the semantic predicate

Etymology information about the historical context (morphological, phonological, syntactic, semantic) of a lexical entry or wordform.

Usage Pragmatic/sociolinguistic information; possible values are:

Region	e.g. dialect
Style	e.g. slang

Frequency corpus-derived frequency of occurrence

4 References

[1] EAGLES/ISLE Metadata Initiative; http://www.mpi.nl/ISLE

[2] Wittenburg, P. (2001) *Lexical Structures*. DOBES internal document. MPI Nijmegen

[3] TEI; http://www.tei-c.org

[4] EAGLES/ISLE; http://www.ilc.pi.cnr.it/EAGLES96

[5] GENELEX; http://www.ilc.pi.cnr.it/EAGLES96/lexarch

[6] SIMPLE; http://www.ilc.pi.cnr.it/EAGLES96

[7] MULTILEX; http://www.ilc.pi.cnr.it/EAGLES96/lexarch

[8] MILE; http://www.ilc.pi.cnr.it/EAGLES96

[9] OLIF; http://www.olif.net

[10] MARTIF; http://www.ttt.org

[11] ISO 12620; http://www.iso.ch/cate/d2517.html

[12] CLS; http://www.ttt.org

[13] SALT; http://www.ttt.org/salt

[14] RDF; http://www.w3.org/RDF

[15] Warwick Framework; http://www.dlib.org/dlib/july96/07weibel.html

[16] Peters, W. (2000) *Metadata for Lexicons*. MPI Lexicon Workshop, Nijmegen

[17] Peters, W. (2001) *Metadata for Lexical Resources*. ISLE CLWG Meeting, Pisa

[18] Gibbon, D. (2001) *Notes on Lexicon Metadata*. ISLE CLWG Meeting, Pisa

[19] Ide, N., Romary, L. (2001) A Flexible Framework for Representing Computational Lexicons. ISLE CLWG Meeting, Pisa

[20] http://www.mpi.nl/DOBES

[21] Bell, J., Bird, St. *A Preliminary Study if the Structure of Lexicon Entries*. http://www.ldc.upenn.edu/exploration/expl2000/program.html

[22] CELEX; http://www.mpi.nl/world/celex

[23] Manning, C. *KirrKirr.* Lexical Tool; http://www.sultry.arts.usyd.edu.au/kirrkirr