



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

D20.2

Specification for Multilingual Medical Dictionary

Report Version: 0.1

Report Preparation Date: 30 June

Classification: RE

Contract Start Date: 2004-01-01

Duration: 3 years

Project Co-ordinator: Hans Åhlfeldt

Department of Biomedical Engineering / Medical Informatics

S-581 83 Linköping University, Sweden

<hans.ahlfeldt@imt.liu.se>



**Project funded by the European Community under the FP6
programme “Integrating and Strengthening the European Research
Area” (2002-2006)**

Table of Contents

1	ADMINISTRATIVE INFORMATION	3
2	INTRODUCTION	3
3	OBJECTIVES OF THE WORK PACKAGE	3
4	INTERCHANGE FORMAT DEFINITION	4
4.1	General considerations	4
4.2	Present step and further developments	4
4.3	List of fields and description	5
4.3.1	Lng	5
4.3.2	Mid	6
4.3.3	Typ	6
4.3.4	Err	6
4.3.5	Lem	6
4.3.6	Mul	7
4.3.7	Frm	7
4.3.8	Mfr	7
4.3.9	Inf	7
4.3.10	Mis	7
4.3.11	Prt	7
4.3.12	Str	8
4.3.13	Ref	8
4.3.14	Exa	8
4.3.15	Com	8
4.4	Basic form	8
4.5	Interchange format	9
4.5.1	MULTEXT definitions	10
4.5.2	XML format for interchange	19
4.5.3	File Delivery Report	21
5	COMMON PLATFORM FOR THE MULTILINGUAL LEXICON	22
5.1	Current State	22
5.2	Outlook	22
6	CONCLUSION	23

1 Administrative information

Lead contractor:	UKLFR (Freiburg University Hospital)
Responsible:	LiU (IMT), DIM, ITRI, JENA, UGOT, SU, INSERM
Participants:	KI, LiU (IDA), CNR, KITH, SOS, STAKES, NBH
Deliverable:	D20.2
Authors:	Stefan Schulz (UKLFR) and Robert Baud (DIM)

2 Introduction

SemanticMining is a EU network of excellence which comprises 25 participant organisations from 11 European countries. This report summarizes the deliverable 20-2 which represents the first version of the Interchange Format Specification for Medical Multilingual Dictionaries which constitutes the main rationale of the work package 20 “Research Activity Multilingual Medical Dictionary”.

The deliverable target audience is the Commission and other project participants. It is not intended for public distribution.

The work package description focuses on multilingual approaches to medical terminology, tackling several aspects such as terminology cross-mapping, cross-language text retrieval, lexicon acquisition and corpus-based machine translation. A central focus of this work package is the construction of a multi-lingual medical base lexicon providing a broad coverage of the main European languages.

3 Objectives of the work package

The work package objectives are to support cross-lingual collaborative terminology development. To this end, it aims at the integration of existing medical lexicons.

- To propose a common data structure for medical lexical resources, linking lexicons of different scope, language and granularity, covering both morphosyntactic and semantic aspects.
- To facilitate the exchange of methods and resources between the groups and to promote joint research.
- To enhance, integrate and populate existing lexicons, using manual, semi-manual and automated lexical acquisition approaches.
- To set up a common platform for the exchange of lexical resources.
- To collect, describe, normalize and exchange domain related corpora to support semi-automated lexeme acquisition.

4 Interchange Format Definition

The following specification is the result of a first draft elaborated by Robert Baud (DIM) in 2004. After its release in December 2004 some WP20 partners used it for converting their lexicons and reported their experiences. During a 2-day workshop in Brighton (Feb 2005) a second version was jointly prepared which was then finished by Robert Baud and released end February. In parallel, a condensed version of this definition was submitted to the 2005 symposium of the American Medical Informatics Association (October 2005). In the meanwhile, the paper was accepted for oral presentation and is currently under revision for final submission.

4.1 General considerations

The Interchange Format is a convention between partners about the way to exchange linguistic information entering in the building process of a medical multilingual lexicon.

The basic idea is that exchange of information is performed through the Interchange Format and only this format. Each contributor is responsible to copy or “translate” his or her data into the Interchange format. On the other end, each receiver is responsible to map the interchange records into his or her own format.

The Interchange format is only for interchange of data, it is not for the design of a central repository, though a number of fields will have exactly the same definition.

Interchange files are available to all partners. No data about initial lexicon entries is available outside of the Interchange files. Subsequent updates are not considered in this statement.

Interchange format is character based with UTF-8 character encoding. This allows 7-bit Ascii to be used transparently, but also supports the full range of Unicode possibilities.

As a consequence, full upper case words are not accepted, except acronyms and a few special situations. The lemmas of lexicon entries have to be specified in their most natural used form, with lower case letters in general and limited occurrences of upper case letters when necessary according to the rules of the language.

In a first step, it is proposed to use a pipe delimited character record, each occurrence of which is a candidate entry for the multilingual lexicon. Such a record is made of a fixed number of fields, either mandatory or optional. The final decision about the acceptance of a candidate is dependent on several criteria: the entry is complete, the entry is valid, the entry is not already present, etc.

The possibility of developing an alternate XML format is left open. The present structure of interchange with a simple list of variables does not require the features of XML and the pipe-delimited format is adequate. See section 4.5 for an example of XML format.

4.2 Present step and further developments

The goal of the interchange format is to create a large corpus of lexicon entries in several languages, typically more than 200'000 to 400'000 entries (though WP20 specifies 50'000 entries only). In this first step, there is no validation between entries and we may be in presence of numerous duplicates (the same entry submitted by different partners). This is not a problem and we will not try to avoid it.

In a second step, we will have to look at each language and to discover the duplicates and carefully eliminate them. The completeness of the delivered entries will be known at this time.

In a third step, the grouping of entries according to distinct object has to be started. There are several methods to do that, either manual and automatic.

Second and third steps are not considered in this documentation.

4.3 List of fields and description

The following list introduces the fields of the interchange format. Each field will be discussed in details in the next sections.

For each field we give a full name, a short name and a definition to be used in the discussion.

Field	Field description	Field definition
Lng	Language	the language to which pertains the present entry, distinguishing Latin and Multilingual (for Proper names)
Mid	Multilingual identifier	the unique identifier of this entry
Typ	Entry type	One of the 4 allowed types of entry
Err	Correctness	flag for correctness of this entry
Lem	Lemma	the entry in its basic form, which will be carefully defined elsewhere
Mul	Syntax argument	the MULTTEXT value of the lemma in basic form
Frm	Inflected form	any inflected form
Mfr	Syntax of inflected form	the MULTTEXT value of the inflected form
Inf	Inflection model	language specific information
Mis	Language specific argument	an argument to be used freely by provider of entries in a specific language
Prt	Decomposition	the decomposition of a compound entry into its parts including infixes or the decomposition of a term into its lexicon entries
Str	Head	the head word of the term
Ref	Referent	the lemma of the referred lexicon entry
Exa	Typical usage	a sentence presenting a typical usage of this entry
Com	Comment	any comment or warning about this entry

4.3.1 Lng

The language field determines to which language this particular entry belongs.

The language field is mandatory from one of the allowed values.

The currently allowed values are: EN for English, FR for French, DE for German, LA for Latin, SV for Swedish, ES for Spanish, PT for Portuguese, -- (double dash) for multilingual. This list will be extended when needed.

The problem of foreign language words has to be considered. For example, if an English word is accepted in a Swedish text, this word is to be considered as a Swedish word and shall duplicate the corresponding English word entry. The two entries are not related one to the other, except that they will probably be pointing to the same Object.

Latin words are a kind of foreign words. As much as possible, they should not be specified in any language except Latin. We want to build a separate Latin section, which can be used in conjunction with any other language. Entering a Latin word under another language raises the risk that this word will appear twice in the executable lexicon. And this is not recommended.

The “Multilingual” language is for words, which are common to all languages, like Proper names. Alzheimer belongs preferably to the “multilingual language” than to English! When a Proper name has a specific value in a given language, it is defined as a new entry for that language.

4.3.2 Mid

This argument acts as a unique identifier of the multilingual lexicon.

The source field is optional.

An interchange file may have a mix of different sources.

A unique identifier of the multilingual lexicon entry, made of the concatenation of two strings separated by a colon: the first string defines the source provider (use your short name as defined in SemanticMining); the second string is a unique identifier within the universe of the source provider (possibly not longer than 10 characters).

4.3.3 Typ

The type of entry defines if we are in presence of a BasicEntry, a SubWordEntry, a CompoundEntry or a TermEntry. By definition, these types are mutually exclusive.

The BasicEntry is for single words of the language, generally without space character in their lemma.

The SubWordEntry is for parts of word entering in the composition of CompoundEntry. A SubWordEntry can generally not be used standalone.

The CompoundEntry is for words, which have been explicitly recognized as a composition of two or more SubWordEntries.

The TermEntry is a sequence of words generally separated by the space character, where the morphology variations of the language may affect any word.

If a provider does not make the distinction between a BasicEntry and a CompoundEntry, he can send BasicEntry only. Further processing in the central repository may be necessary to resolve the CompoundEntries.

The type of entry field is mandatory from one of the four possible values. The allowed values are: B for BasicEntry, S for SubWordEntry, C for CompoundEntry and T for TermEntry.

4.3.4 Err

The correctness is the quality of an entry of being valid in its language. Incorrect words, which are frequently seen, may be part of the lexicon, but they are marked as such.

This variable is optional. An empty value means a correct entry. The following values are accepted: M for mistyping error, F for orthographic error, S for simplification of the language, X for other errors.

4.3.5 Lem

The lemma is the representation of the entry in its basic form, as an ordered set of printable characters. It is a string of any length.

The lemma field is mandatory.

The basic form is defined elsewhere in this document. It is supposed to be derivable from any occurring form, using a flexional morphology process, which is language dependant. There is a unique basic form for any entry.

4.3.6 Mul

The syntax argument is defined by the open standard MULTEXT. Language dependent extensions of MULTEXT may be used.

This argument is generally mandatory. It should be as complete as possible for the several allowable subfields. However, in special situations like Latin and Proper names, this argument may be empty.

It has been decided to conform to MULTEXT for this field and to use its subcategory fields for specific variations necessary for the multilingual lexicon.

The following values are considered as an extension of MULTEXT: F for subwords and other part of words entering in the composition of a CompoundEntry and known as SubWordEntries. A type subcategory specifies the position: p for prefix, s for suffix.

4.3.7 Frm

Any inflected form of interest may be given in this field.

This argument is optional.

The idea behind this argument is to allow to export all the inflected forms of a lexicon entry on an automated basis. Such a function depends on the existence of a language-specific generator of inflected forms.

4.3.8 Mfr

The syntax argument of the inflected form, using MULTEXT exactly as for the Mul argument.

This argument is mandatory if the Frm field is present, empty otherwise.

4.3.9 Inf

This argument is language-specific and may be used by the provider to specify to which kind of inflection the present entry belongs.

This argument is optional.

4.3.10 Mis

This argument is to be understood as a catch-all argument at disposal of the provider for additional information of its own.

This argument is optional.

Any provider should be aware that overuse of this argument will bring a problem of integration at some future time. As a general statement, when one observes a trend of systematic use of this field, the provider should ask for a revision of the present specifications.

4.3.11 Prt

The decomposition holds for CompoundEntry and provides an ordered set of all parts entering in the composition of a CompoundEntry.

The decomposition field is optional. It can be automatically reconstructed later.

In principle, this field is generally made of existing SubWordEntry and occasionally BasicEntry. But this constraint will not necessarily be verified, when feeding the lexicon with new words.

For the CompoundEntry, the decomposition field is made of all the component entries separated by a double underscore. Between the underscores, a possible infix or a dash can take place. When removing the underscores and only them, we should find the lemma.

4.3.12 Str

The head entry is the dominant part in a TermEntry, starting from which the term is retrieved in a sentence. It is not necessarily the first part.

The head field is mandatory for entries of the TermEntry type.

4.3.13 Ref

The referent is the lemma value of another entry to which the present entry makes reference. Alternatively and preferably, this variable can be a Mid.

The referent field is mandatory for acronyms, abbreviations and mistypings.

This field should be filled in independently of the existence of the referred entry. No check will be done about this fact.

We have to recognize a potential problem with this field because the lemma does not uniquely identifies a lexicon entry. The solution is outside of the scope of this interchange format, when creating a unique identifier for each entry. However, the extension of this problem is quite limited.

4.3.14 Exa

One typical sentence showing the use of the current entry. This variable is just here for documentation purpose.

This field is optional.

4.3.15 Com

The comment is the place for additional short information about unexpected situation. This field should not be overused (< 5%).

The comment field is optional.

4.4 Basic form

The basic form is one of the inflectional morphology variants of any lexicon entry. It is selected for its convenience according to common practice. The basic form being language dependent, it is recommended to follow the most common practice of the reference dictionaries of that language.

The basic form is often defined as:

the singular number,

the masculine gender,

the nominative case,

the infinitive mode for verbs,

whatever applies, depending on the part of speech and the given language.

For some lexicon entries, it is known that the recommended basic form as above does not exist. In this situation the closest existing representation is selected, whatever it is.

The basic form for a CompoundEntry is generally obtained by the selection of the basic form of its last part, the other parts being naturally in their basic form.

The basic form of a TermEntry is obtained by replacement of the head of the term by its basic form, the rest of the term being left as it is.

4.5 Interchange format

The interchange format is the following:

Text file with UTF-8 characters.

1 entry per line (each line is terminated by the characters #13 and #10 or #13 only).

A line is an ordered sequence of a fixed number N of fields.

The N fields are separated (N – 1 times) by the separator character |.

Non terminal empty fields result in two consecutive separator characters.

At the time of publication of the present document, N has the value 15 and the order is defined above (see table on the second page)

4.5.1 MULTEXT definitions

(The following text is an extraction from MULTEXT).

4.5.1.1 Attribute/value tables

The categories listed below with the relevant attributes and values are based on EAGLES documents and are the results of a first testing based on a proposal made by Veronis et al. 1994 for lexical specifications in MULTEXT. As it has already been mentioned in the section "Background considerations" that propose features for describing lexical items of different languages aiming at defining a set which can be said "common" for all of them is a complex task. The underlying philosophy for this task has then be to lead different groups into a pragmatic solution where the concept of an "harmonized" set of features could be reached.

The groups have first worked out their lexical descriptions taking as input EAGLES and Veronis et al. (1994) documents. The very general criterion was to encode those proposed features which were considered relevant for the language in question. Therefore MULTEXT also followed EAGLES bottom-up methodology in trying to define extensively the features "used" in the lexical descriptions for each group language, as this procedure will make evident the features commonly used. After this phase, whose result can now be seen in the section "Comparison of attribute/values used by the groups", a new phase is envisaged as to accommodate language-specific considerations into a general model to be used by MULTEXT. This accommodation must take into account extensibility to other languages and also application motivated arguments, as well as internal coherence. For this new phase more specific criteria would be desirable with respect the addition of new features to the EAGLES Level-1 set. The aimed result is a "harmonized" set of features which properly describe lexical items of the different languages.

Following the general aim of the project, these harmonized specifications -- and the related resources -- will contribute to the standardization of the corpus annotation work. They are supposed to serve as a user oriented additional characteristic of our tool package in the sense that end-users will have a common ground for inspecting and understanding the resources and tool results independently to a large extent of the language. This common set of features will also be a common ground to perform comparisons of different annotation tool results, because, as mentioned in the previous section, the existence of many lexical description systems is causing nowadays a problem for comparing results. Therefore the categories and features listed below are the common reference for the work done by the different groups. Further discussion on this first proposal is to be found in the section "Comparison of the attributes/values used by the groups" which is in turn to define criteria for changing this first proposal.

4.5.1.1.1 Part of Speech Table

```

=====
Part-of-Speech Code
=====
Noun           N
Verb           V
Adjective      A
Pronoun        P
Determiner     D (for those who do not have a separate category)
Article        T (for Articles, these are included in Determiner)
Adverb         R
Adposition     S
Conjunction    C
Numerals       M
Interjection   I
Unique         U
Residual       X
Abbreviation   Y
=====

```

Each character at positions 1, 2, etc. encodes the value of one attribute (person, gender, number, etc.), according to the tables given below.

Abbreviations used:

P Position (starts with 0 for encoding PoS values)
 ATT Attribute name
 VAL Value
 C Code

4.5.1.1.2 Nouns (N)

```

= =====
P  ATT      VAL      C
= =====
1  Type     common   c
   Type     proper   p
-----
2  Gender   masculine m
   Gender   feminine f
   Gender   neuter   n
-----
3  Number   singular s
   Number   plural   p
-----
4  Case     nominative n
   Case     genitive  g
   Case     dative   d
   Case     accusative a
= =====

```

4.5.1.1.3 Verbs (V)

P	ATT	VAL	C
1	Type	main auxiliary modal	m a o
2	Mood/VForm	indicative subjunctive imperative conditional infinitive participle gerund supine base	i s m c n p g s b
3	Tense	present imperfect future past	p i f s
4	Person	first second third	1 2 3
5	Number	singular plural	s p
6	Gender	masculine feminine neuter	m f n

4.5.1.1.4 Adjectives (A)

=	=====	=====	=====
P	ATT	VAL	C
=	=====	=====	=====
1	Type	qualificative ordinal cardinal indefinite possessive	f o c i s

2	Degree	positive comparative superlative	p c s

3	Gender	masculine feminine neuter	m f n

4	Number	singular plural	s p

5	Case	nominative genitive dative accusative	n g d a
=	=====	=====	=====

4.5.1.1.5 Pronouns (P)

= =====	=====	=====
P ATT	VAL	C
= =====	=====	=====
1	Type	personal p demonstrative d indefinite i possessive s interrogative t relative r exclamative e reflexive x reciprocal l
- - - - -		
2	Person	first 1 second 2 third 3
- - - - -		
3	Gender	masculine m feminine f neuter n
- - - - -		
4	Number	singular s plural p
- - - - -		
5	Case	nominative n genitive g dative d accusative a oblique o object j
- - - - -		
6	Possessor	singular s plural p
= =====	=====	=====

4.5.1.1.6 Determiners (D)

P	ATT	VAL	C
=====			
1	Type	demonstrative indefinite possessive interrogative	d i s t

2	Person	first second third	1 2 3

3	Gender	masculine feminine neuter	m f n

4	Number	singular plural	s p

5	Case	nominative genitive dative accusative oblique	n g d a o

6	Possessor	singular plural	s p
=====			

4.5.1.1.7 Articles (T)

P	ATT	VAL	C
1	Type	definite indefinite	d i
2	Gender	masculine feminine neuter	m f n
3	Number	singular plural	s p
4	Case	nominative genitive dative accusative	n g d a

4.5.1.1.8 Adverbs (R)

P	ATT	VAL	C
1	Type	general particle	g p
2	Degree	positive comparative superlative	p c s

4.5.1.1.9 Adpositions (S)

P	ATT	VAL	C
1	Type	preposition postposition circumposition	p t c
2	Formation	simple compound	s c

4.5.1.1.10 Conjunctions (C)

P	ATT	VAL	C
1	Type	coordinating subordinating	c s

4.5.1.1.11 Numerals (M)

=	=====	=====	=====
P	ATT	VAL	C
=	=====	=====	=====
1	Type	cardinal	c
		ordinal	o
-	-----	-----	-----
2	Gender	masculine	m
		feminine	f
		neuter	n
-	-----	-----	-----
3	Number	singular	s
		plural	p
-	-----	-----	-----
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
=	=====	=====	=====

4.5.1.1.12 Interjections (I)

4.5.1.1.13 Unique membership class (U)

4.5.1.1.14 Residual (X)

4.5.1.1.15 Abbreviations (Y)

4.5.2 XML format for interchange

The interchange format has been selected as a pipe-delimited format on the basis that it has only to vehicle a single list of variables without structure. However, the XML standard, though not strictly necessary in the present situation, has to be considered as an alternative solution because of its universal acceptance.

We here demonstrate how the actual pipe-delimited file can be automatically converted to an XML structure and vice-versa. This opportunity is uniquely presented here for documentation purpose and no decision about using the XML format has been taken.

The text below represent a sample file using the Interchange format:

```
lng|Mid|Typ|Err|Lem|Mul|Frm|Mfr|Inf|Mis|Prt|Str|Ref|Exa|Com
FR|HUG:123456|B||humérus|cmsn|humérus|cmpn||||fracture de l'humérus gauche|
FR|HUG:123457|B||huméral|fpmsn|||||plaie humérale|
```

It can be seen that a first line contains the list of all fields by name, followed by 2 entries.

The equivalent file in XML format can be obtained automatically and has the present layout:

```
<?xml version="1.0" encoding="iso-8859-1" standalone="yes" ?>
- <FILE TABLE="sample interchange format">
  - <FIELD COUNT="15">
    <COLUMN POS="1">lng</COLUMN>
    <COLUMN POS="2">Mid</COLUMN>
    <COLUMN POS="3">Typ</COLUMN>
    <COLUMN POS="4">Err</COLUMN>
    <COLUMN POS="5">Lem</COLUMN>
    <COLUMN POS="6">Mul</COLUMN>
    <COLUMN POS="7">Frm</COLUMN>
    <COLUMN POS="8">Mfr</COLUMN>
    <COLUMN POS="9">Inf</COLUMN>
    <COLUMN POS="10">Mis</COLUMN>
    <COLUMN POS="11">Prt</COLUMN>
    <COLUMN POS="12">Str</COLUMN>
    <COLUMN POS="13">Ref</COLUMN>
    <COLUMN POS="14">Exa</COLUMN>
    <COLUMN POS="15">Com</COLUMN>
  </FIELD>
  - <REC NO="1">
    <lng>FR</lng>
    <Mid>HUG:123456</Mid>
    <Typ>B</Typ>
    <Err />
    <Lem>humérus</Lem>
    <Mul>cmsn</Mul>
    <Frm>humérus</Frm>
    <Mfr>cmpn</Mfr>
    <Inf />
    <Mis />
    <Prt />
    <Str />
    <Ref />
    <Exa>fracture de l'humérus gauche</Exa>
    <Com>invariable at plural</Com>
  </REC>
  - <REC NO="2">
    <lng>FR</lng>
    <Mid>HUG:123457</Mid>
    <Typ>B</Typ>
    <Err />
    <Lem>huméral</Lem>
    <Mul>fpmsn</Mul>
    <Frm />
    <Mfr />
    <Inf />
    <Mis />
    <Prt />
    <Str />
    <Ref />
    <Exa>plaie humérale</Exa>
    <Com />
  </REC>
  <TOTAL>2</TOTAL>
  <DATE>01.06.2005</DATE>
</FILE>
```

A reverse processing of the XML file is able to rebuild the initial pipe delimited file at any time.

4.5.3 File Delivery Report

This document should be instantiated as an electronic submission form, which acts as an accompanying document with each interchange file. It is supposed to be transposed in an XML document for permanent storage. In a final implementation the submission form will have to be fill in on an internet server and the XML document will be automatically generated.

For every submission of a lot of lexicon entries using the Interchange Format, the provider is invited to fill in the File Delivery Report. Such a document specifies or confirm a few parameters, which may be necessary for an adequate processing of the submitted lexicon entries.

The following is a representation (not its implementation) of the content of the File Delivery Report.

Source identification
Source acronym: _____
Provider name: _____
Date of delivery: __.__.____ (dd.mm.yyyy)
Content description
Number of entries, exact: _____ estimated: _____
Language: __ (one of EN, FR, DE, SV, LA, ES, PO, IT, --)
Quality and comments: _____
Transmission information
File name: _____ (first 3 letters of acronym + ... + .sem)
Separator character: __ (default is)
Line termination: CRLF or CR (select one option)
File size: _____ Kb (not compressed)
Found information before insertion
Date of processing: __.__.____ (dd.mm.yyyy)
Number of entries: _____
Validated entries: _____ Rejected entries: _____
Main reason for rejections: _____
Found information after insertion
Date of processing: __.__.____ (dd.mm.yyyy)
By whom: _____
Number of entries: _____
Validated entries: _____ Rejected entries: _____
Main reason for rejections: _____

(the electronic version of this report is to be later implemented)

5 Common Platform for the Multilingual Lexicon

5.1 Current State

We are using the MERMIG content management system as a common platform for uploading the converted lexicons. Currently there are five different sources totaling 120,666 lexemes (cf. Figure 1).



The screenshot shows the MERMIG content management system interface. At the top, there are navigation tabs: WEB MANAGER, DOCUMENT MANAGER, GROUP MANAGER, CALENDAR, FORUM, E-MAIL/SMS, and SEARCH. Below these is a 'WORK GROUP' dropdown menu with options: WP20 - Lexicon, WP1, WP2 - Planning, WP3 - Common DB, WP4, and WP5. A breadcrumb trail reads 'Document Manager > Top/ Common Lexicons & Corpora/ Lexicons'. Below the navigation is an 'Abstract' section with a search bar: 'List items containing * in Any Field'. The main content is a table with columns: Check all, Title, Items, Owner, Size, and Date +.

Check all	Title	Items	Owner	Size	Date +
	Previous section				
	Lexicons - Previous Versions	0			
	<input type="checkbox"/> Swedish lexicon		dimitrios kokkinakis	208.83 KB	2005-06-14
	<input type="checkbox"/> Swedish lexicon - File Delivery report		dimitrios kokkinakis	17.175 KB	2005-06-14
	<input type="checkbox"/> German Lexicon		Susanne Hanser	2440.763 KB	2005-06-17
	<input type="checkbox"/> German Lexicon File Delivery Report		Susanne Hanser	10.837 KB	2005-06-17
	<input type="checkbox"/> Morphosaurus Lexicon File Delivery Report	new	Kornél Markó	7.78 KB	2005-06-20
	<input type="checkbox"/> Morphosaurus Lexicon	new	Kornél Markó	2898.204 KB	2005-06-20
	<input type="checkbox"/> English lexicon from LIU	new	Magnus Merkel	48.273 KB	2005-06-23
	<input type="checkbox"/> Swedish lexicon from LIU	new	Magnus Merkel	37.889 KB	2005-06-23
	<input type="checkbox"/> File delivery report (LIU English lexicon)	new	Magnus Merkel	23.709 KB	2005-06-23
	<input type="checkbox"/> File delivery report (LIU Swedish lexicon)	new	Magnus Merkel	23.697 KB	2005-06-23

Figure 1: Repository for Lexicon Uploads

These sources constitute a merger of heterogeneous medical dictionaries produced at different locations (Freiburg, Göteborg, Linköping, Geneva, Paris) covering medical terminology of several languages (English, Swedish, German, French, Spanish, Portuguese). The lexicons have different specifications (words, multiword entries, subword entries) and different levels of coverage, granularity, correctness and completeness. As all these lexicons are still under development, and periodic updates will be released.

5.2 Outlook

In the framework of the research activities within the NoE SemanticMining the common pool of lexical sources constitutes an important starting point for joint research in lexicon acquisition and semi-automated term translation.

Next steps will further include:

- Generation and upload of fully-formed lexicons (with contain all inflectional variants)
- Update of the lexicon specification in the light of new experiences and new (external) standards
- Enhancement of the specification by including basic semantic data (source-dependent synonyms, homonyms, translations)
- Inclusion of additional sources, e.g. the English UMLS Specialist lexicon.

- Setup of a corpus repository, due to the high importance of unrelated, related and parallel corpora for automated lexicography.

6 Conclusion

The common interchange format constitutes an important interfacing activity which brings together lexical data from different locations, different languages, as well as diverging granularities and focuses. The WP20 partners which have committed themselves to put this into effect have long experiences with medical linguistics and lexicography. It is therefore realistic to reach the goal to build up the world wide biggest repository of medical language resources by the end of the year 2006.