



Methodology and Architecture for Multilingual Resources Management

Distribution: Project internal

MedIEQ Quality Labeling of Medical Web content using Multilingual Information Extraction

National Centre for Scientific Research "Demokritos"
Teknillinen Korkeakoulu – Helsinki University of Technology
Universidad Nacional de Educación a Distancia
Col·legi Oficial de Metges de Barcelona
Zentralstelle der deutschen Ärzteschaft zur Qualitätssicherung in der
Medizin
Vysoka Skola Ekonomicka v Praze
I-Sieve Technologies Ltd

2005107 **D10**

December 2006

| | |
|--------------------|--|
| Project ref. no. | <i>2005107</i> |
| Project acronym | <i>MedIEQ</i> |
| Project full title | <i>Quality Labeling of Medical Web content using Multilingual Information Extraction</i> |

| | |
|------------------------------------|---|
| Security (distribution level) | <i>Project internal</i> |
| Contractual date of delivery | <i>30 September 2006</i> |
| Actual date of delivery | <i>28 December 2006</i> |
| Deliverable number | <i>D10</i> |
| Deliverable name | <i>Methodology and Architecture Multilingual Resources Management</i> |
| Type | <i>Report</i> |
| Status & version | <i>Final</i> |
| Number of pages | <i>22</i> |
| WP contributing to the deliverable | <i>WP7</i> |
| WP / Task responsible | <i>UNED</i> |
| Other contributors | <i>NCSR, UEP, TKK, WMA, AQUMED</i> |
| Author(s) | <i>E. Amigó, C. Muñoz, F. López-Ostener, V. Peinado (UNED)</i> |
| EC Project Officers | <i>Artur Furtado, Zinta Podniece</i> |
| Keywords | <i>lexical and semantic resources, repository, browsing</i> |
| Abstract (for dissemination) | <p><i>This document proposes the methodology and architecture for multilingual resources management in MedIEQ (as described in the Technical Annex).</i></p> <p><i>MRM will be a set of tools/components for the management of linguistic resources in different languages. Components take care of operations with resources: add, remove, load, select, extract, etc.</i></p> |

Table of contents

| | |
|--|----|
| Table of contents | 4 |
| 1. Introduction | 5 |
| 2. Starting Point: Conclusions after Barcelona Meeting | 5 |
| 3. Resources | 6 |
| 3.1 Resources Questionnaires..... | 6 |
| 3.2 First Pool of Resources Proposed | 6 |
| 3.2.1 Format and Encoding Requirements | 6 |
| 3.3 Definitive Resources: UMLS & UMLSKS | 7 |
| 4. Components, Architecture and Functionalities..... | 8 |
| 4.1 Description of MRM Components..... | 9 |
| Resources Manager..... | 9 |
| Repository..... | 9 |
| Generic Parser | 9 |
| Browser | 10 |
| Converter..... | 10 |
| 4.2 Functionalities of the MRM Components..... | 10 |
| 4.3. Modification after the Helsinki Meeting..... | 11 |
| 5. Use Cases | 12 |
| Use case 1: Add a resource..... | 12 |
| Use case 2: Remove a resource..... | 12 |
| Use case 3: Load a resource | 12 |
| Use case 4: Select elements/parts of a resource | 13 |
| Use case 5: Extract the selected elements/parts of a resource | 13 |
| Use case 6: Translate the selected elements/parts of a resource | 13 |
| Use case 7: Create a custom resource | 13 |
| Use case 8: Modify format/encoding of an existing resource | 14 |
| Use case 9: Manipulate the available formats | 14 |
| Use case 10: Manipulate the available encodings | 14 |
| Use case 11: Manipulate the available languages | 14 |
| 6. Interfaces..... | 14 |
| 6.1 Expert Interface | 15 |
| 6.2 SysAdmin Interface | 15 |
| 7. Design and Work Plan | 15 |
| 7.1 Capabilities to be Implemented in the 1 st Version of the MRM Toolkit | 15 |
| 7.2 Work Assignments and Future Plans..... | 16 |
| 8. Concluding Remarks | 16 |
| APPENDIX A: Further Details about Key Resources and Tools | 17 |
| Eurovoc..... | 17 |
| ICD-10..... | 17 |
| MeSH..... | 18 |
| MetaMap | 18 |
| MMTx..... | 19 |
| MTI..... | 19 |
| SNOMED | 20 |
| UMLS | 20 |
| UMLSKS | 21 |

1. Introduction

MRM will be a set of tools/components for the management of linguistic resources in different languages. It will be accessed by two different users and it will therefore provide two user interfaces (UIs).

On one hand, the toolkit will allow to import linguistic and semantic resources in pre-defined supported formats into a repository and to access them in order to manipulate their contents. Once a given resource is imported into a repository, it will be possible to browse across its contents and to create new resources by extracting and merging data.

On the other, the toolkit will integrate some Natural Language Processing (NLP) tools in order to e.g. process some input text and to automatically suggest terms and descriptors from a controlled vocabulary. All these tools are already available in the proposed resources packages. Therefore, in this case, the MRM Toolkit will work as an extra-layer in top of the available resources and tools in order to allow communication between the resources repositories and the other components.

Finally, the description of the work done in this document is chronological so that the proposed ideas, the modifications and the feedback received after the meetings with other partners and the Advisory Committee members are included.

2. Starting Point: Conclusions after Barcelona Meeting

After the Barcelona Meeting in June 1-2, 2006, two main conclusions were taken affecting the development of the activities proposed in the WP7 technical annex¹:

- The MRM will work in a multilingual environment but will not need cross-language capabilities. Therefore, it will contain resources in every language involved in the project (in the early steps, only English and Spanish are considered) but not any tools for translation.
- At least for its early versions, the MRM Toolkit will not incorporate Natural Language Processing tools and will focus on the addition, manipulation and maintenance of linguistic resources, specifically, controlled vocabularies over health domain.

¹ The minutes of the Barcelona meeting are available at:
http://www.medieq.org/system/files?file=MedIEQ_2ndmeetingminutes_June1-2-2006_Barcelona.doc

3. Resources

3.1 Resources Questionnaires

UNED prepared a set of questionnaires and asked the other partners to provide information about the linguistic resources and NLP tools owned by each group, their availability and cost, in order to evaluate the possible usefulness for the MedIEQ project.

However, most of these resources were discarded, for the time being, because:

- They were available in languages not to be considered at the first stages of the projects (Greek, Czech, Finnish, etc.).
- They were either general purpose or domain-specific but in both cases it was difficult to apply them on the medical domain (newswire collections, financial or technical corpora, etc.).
- They were language-independent tools based on Machine Learning techniques which cannot be effectively used without the appropriate training data (POS taggers), which we do not have.

3.2 First Pool of Resources Proposed

Multilingual thesauri, lexica, vocabularies, etc. have to be effectively handled. Among the linguistic resources we originally proposed, we can find:

- Medical Subject Headings² (MeSH) and its versions in other languages.
- International Classification of Diseases³ (ICD).
- Eurovoc⁴: multilingual, polythematic thesaurus focusing on the law and legislation of the European Union and covering domains of interest for the European Union.

3.2.1 Format and Encoding Requirements

Usually, such resources are provided in XML format, but other known formats have also to be considered, namely:

- XML/RDF
- semi-structured ASCII
- word lists, comma-separated values (CSV)

2 See Appendix A for further details about MeSH.

3 See Appendix A for further details about ICD.

4 See Appendix A for further details about Eurovoc.

Internally, the MRM Toolkit will work in Unicode (utf-8) but other encodings will be supporting for importing/exporting purposes. The encodings to be considered in the preliminary version (English and Spanish, only) are:

- Unicode (utf-8)
- iso-8859-1/iso-8859-15
- us-ascii
- windows-1252

Additional encodings will be considered in the final version of the toolkit when other languages become available, e.g.:

- windows-1250 (Czech)
- windows-1253 (Greek)

3.3 Definitive Resources: UMLS & UMLSKS

Once we have decided to focus on the MeSH controlled vocabulary and its versions in other languages, all partners agreed to use the Unified Medical Language System (UMLS)⁵ components as main resource, specifically, the UMLS Metathesaurus.

The main advantages of this decision are:

- it is an effective way of obtaining/updating the resources, since UMLS is distributed as a unique package.
- it already contains the standard version of MeSH hierarchy in most of the languages involved in the MedIEQ project.
- for compatibility reasons, it provides the essential mapping between different controlled vocabularies, e.g. MeSH↔SNOMED-CT, MeSH↔ICD.
- it contains additional resources/tools we may easily add in the future.

In order to access, manipulate and integrate all the resources available in UMLS within the MRM toolkit, we will also use the UMLSKS services, specifically the JAVA API provided for developing purposes. Thus, the MRM Toolkit will work as an extra-layer in top of the UMLS's datasets and tools in order to allow communication between the resources repositories and the other components.

In addition, other UMLSKS tool proposed to be integrated within the MRM toolkit is MMTx⁶. MMTx is an effort to make the MetaMap program available to biomedical researchers in a generic, configurable environment. MetaMap maps arbitrary text to concepts in the UMLS Metathesaurus; or, equivalently, it discovers Metathesaurus concepts in

5 See Appendix A for further details about UMLS, its components and the UMLSKS tools and APIs.

6 Again, see Appendix A for further information about MMTx and MetaMap.

text. These capabilities may be convenient and useful for our experts during the accreditation process of medical content.

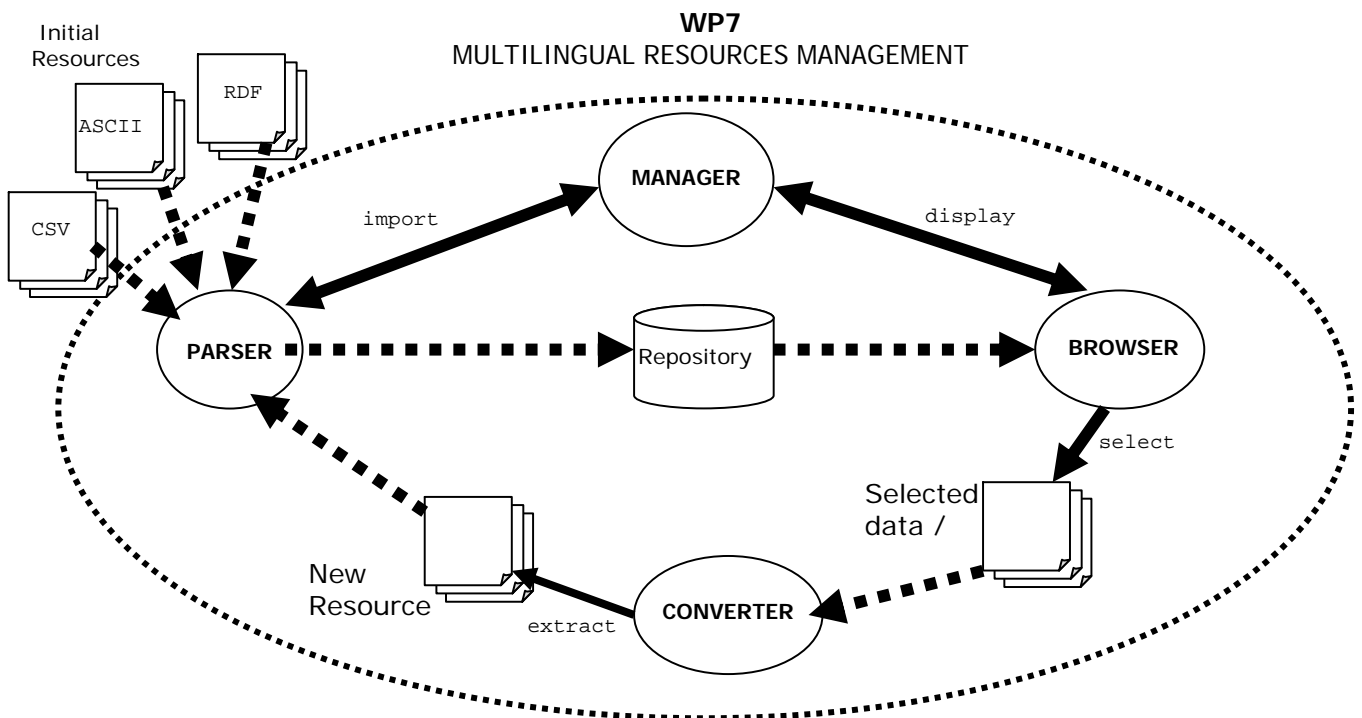
However, there are some open issues that need agreement with other technical partners:

- UMLS resources can be easily transformed into databases both MySQL and Oracle formats and the Aqua's databases are built upon a Postgresql server. Even though there are tools allowing conversion among these formats, no tests has been performed yet.
- It is necessary to test the feasibility of the database schema provided by UMLS within the general database design of the Aqua system.

4. Components, Architecture and Functionalities

A set of tools/components will constitute the Multilingual Resources Management toolkit. The toolkit will provide user interfaces through which several actions will be possible.

In order to have a better image on its architecture, see schema 1 below.



Schema 1: WP7 Toolkit & Architecture (preliminary version)

4.1 Description of MRM Components

A component tree of the MRM toolkit is shown below.

```
Multilingual Resources Manager (MRM)
|
|--- Resources Manager
|
|--- Repository
|
|--- Generic Parser
|   |-- Specific Parsers
|
|--- Browser
|
|--- Converter
```

Resources Manager

Main interface showing the general options and allowing access to the rest of the components within the toolkit in order to manipulate the Repository of resources.

Repository

Local database containing the original copy of the resources included in the toolkit and their internal representation. Besides, some other metadata about the resources themselves will be stored such as type of resources, available languages, etc.

Custom-made resources will also be included in the Repository, along with their associated metadata (e.g. original sources, date of creation, creator, language, etc).

Generic Parser

Takes as input one of the supported resources, analyzes it and creates a data representation in the Repository which allows Browser and Converter to manipulate the data.

Each resource can be added from sources in different format, namely: XML/RDF, ASCII, CSV, etc. To solve this problem, specific Parsers will be designed in order to read the input format and create a common (or at least compatible) data representation in the Repository.

Specific parsers may also be employed when parsing custom resources.

Browser

Loads a selected resource and allows browsing across its structure. When data are clearly structured, `Browser` visualizes them as a tree. When data are made of plain structures, it shows them as a table or as a simple list of terms.

Converter

Takes as input the data extracted with the `Browser` and creates a new resource or changes the format and/or the encoding of the resource.

4.2 Functionalities of the MRM Components

Resources Manager

Generic Parser

The functionalities of this component are:

- a) Parse a resource
- b) Identify the input encoding and transform it into utf-8, whenever it's necessary
- c) Save a local copy of the original version of the resource

Parser: identifies the input encoding, transforms it into utf-8 when necessary, parses the contents and saves a local copy of the resource.

Browser

The functionalities of this component are:

- a) Load a resource included in the Repository
- b) Select (a part of) the structure/some elements of the resource
- c) Extract (a part of) the structure/some elements of the resource
- d) Cross-Language browsing: allows to jump from a resource in a given language to its counterpart in another available language, e.g.: moving from a subtree of the English MeSH to its counterpart into the Spanish MeSH.

Converter

The functionalities of this component are:

- a) Export (a part of) the structure/some elements/sub-trees into a new resource
- b) Change format
- c) Change encoding

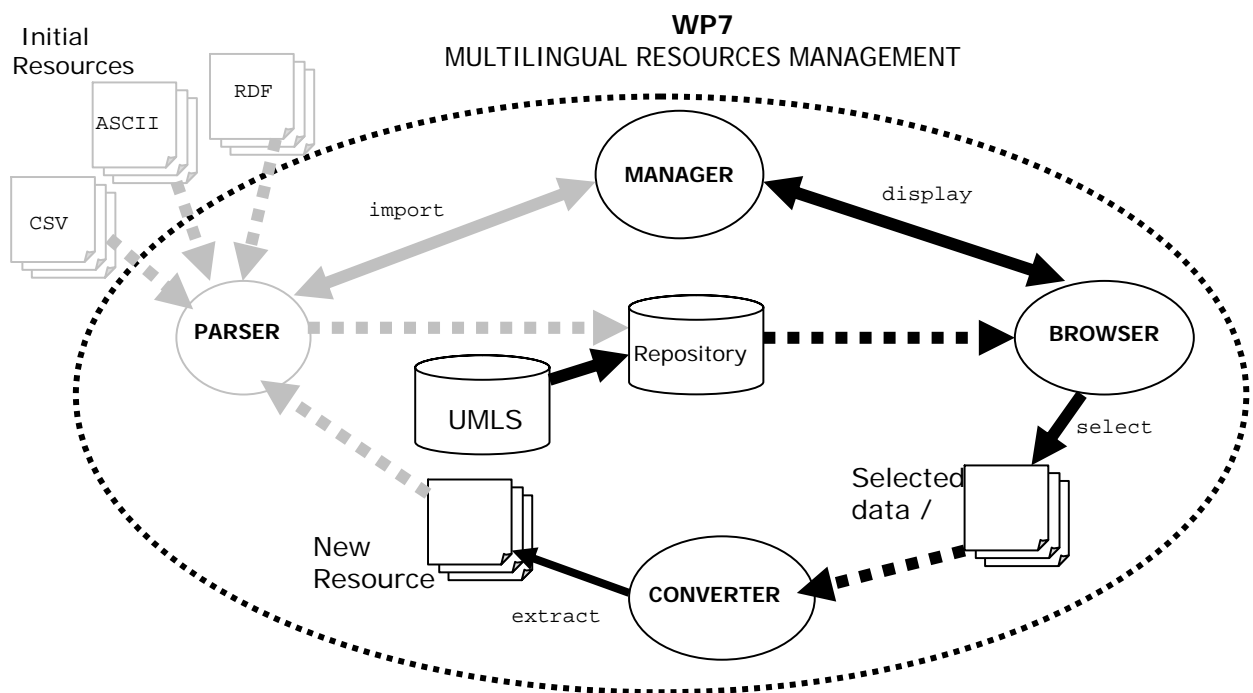
4.3. Modification after the Helsinki Meeting

After the Helsinki Meeting in October 23-24, 2006, some decisions were taken that force to modify the initial proposal.

First, by the time being, the MRM toolkit will only handle resources included in UMLS, so the development of the functionalities related to parsing new resources, adding resources into the repository, managing input formats, etc. are temporarily stopped, expecting future decisions.

Dr. Stéphan Darmoni, member of the Advisory Committee, recommended to integrate, along with MeSH, the SNOMED-CT⁷ medical terminology, also included in UMLS. Dr. Darmoni also recommended to test Medical Text Indexer (MTI) as an alternative method of discovering MeSH headings for citation titles and abstracts and suggesting indexing terms⁸.

Schema 2 shows the updated version of Architecture after the Helsinki meeting.



Schema 2: WP7 Toolkit & Architecture (updated version after the Helsinki meeting)

⁷ See Appendix A for further details about SNOMED-CT.

⁸ Again, see Appendix A for a more comprehensive description of MTI.

5. Use Cases

Use case 1: Add a resource

Actors

[-H-] Labelling expert, Resources Manager

[-L-] Parser, Repository

Interactions

[-H-] A labelling expert asks the Resources Manager to add a resource by selecting the path to the given resource. The latter must have one of the accepted formats.

[-L-] The Parser analyzes the resource's structure and content and adds it into the Repository.

Use case 2: Remove a resource

Actors

[-H-] Labelling expert, Resources Manager

[-L-] Repository

Interactions

[-H-] A labelling expert selects a given resource to be removed.

[-L-] The selected resource is deleted.

Use case 3: Load a resource

Actors

[-H-] Labelling expert, Browser

[-L-] Repository

Interactions

[-H-] A labelling expert selects a given resource from the Repository to be accessed.

[-L-] The Browser loads and displays it.

Use case 4: Select elements/parts of a resource**Actors**

[-H-] Labelling expert, Browser

Interactions

[-H-] A labelling expert selects elements/parts of a loaded resource in order to, e.g., extract/translate them.

Use case 5: Extract the selected elements/parts of a resource**Actors**

[-H-] Labelling expert, Browser

Interactions

[-H-] A labelling expert asks the Browser to extract the selected elements/parts of a loaded resource in order to, e.g., create a new resource.

Use case 6: Translate the selected elements/parts of a resource**Actors**

[-H-] Labelling expert, Browser

[-L-] Repository

Interactions

[-H-] A labelling expert asks the Browser to translate the selected elements/parts of a loaded resource into other available language.

[-L-] The Browser loads the resource's counterpart from the Repository to allow the translation/movement between languages.

Use case 7: Create a custom resource**Actors**

[-H-] Labelling expert, Converter

Interactions

[-H-] A labelling expert asks the Converter to create a custom resource by, e.g., merging parts of existing resources and/or adding items.

Use case 8: Modify format/encoding of an existing resource**Actors**

[-H-] Labelling expert, Converter

Interactions

[-H-] A labelling expert asks the Converter to change the format/encoding of an existing resource.

Use case 9: Manipulate the available formats**Actors**

[-H-] SysAdmin, Resources Manager

Interactions

[-H-] Add/remove/change a given format to make it available.

Use case 10: Manipulate the available encodings**Actors**

[-H-] SysAdmin, Resources Manager

Interactions

[-H-] Add/remove/change a given resource encoding to make it available.

Use case 11: Manipulate the available languages**Actors**

[-H-] SysAdmin, Resources Manager

Interactions

[-H-] Add/remove/change a given language to make it available.

6. Interfaces

Two different UIs should be available in each tool/component: one for the expert user (specialists from AQUED & WMA) and a second for the system administrator.

6.1 Expert Interface

The Resources Manager will show the general options and access to the rest of the components within the toolkit in order to manipulate the Repository of resources. The actions implemented in this component are:

- a) Import: add a new resource to the Repository (through the Parser) when this resource is available in one of the formats supported by the MRM Toolkit. This action takes place only when including a new resource or a new version of a resource.
- b) Remove: delete a resource from the Repository
- c) Display: visualize/browse (through the Browser) a resource available in the Repository in a convenient way
- d) Create a new resource from existing resources in the Repository (through the Converter)

6.2 SysAdmin Interface

The SysAdmin interface will allow enabling/disabling the functionalities implemented in the toolkit. Most of them will appear as options in a multi-check list or combo boxes.

- a) Add/remove/modify the supported formats.
- b) Add/remove/modify the supported encodings
- c) Add/remove/modify the supported languages.

7. Design and Work Plan

7.1 Capabilities to be Implemented in the 1st Version of the MRM Toolkit

The capabilities implemented in the 1st version of the MRM Toolkit (due to M12, according to the Technical Annex) are the following:

On one hand, the Expert UI will allow to:

- Use case 3: Load a resource from the Repository into the Browser
- Use case 4: Select elements/parts of a resource
- Use case 5: Extract the selected elements/parts of a resource
- Use case 7: Create a custom (un-structured) resource
- Use case 8: Modify format/encoding of an existing resource

On the other hand, the SysAdmin UI will allow:

- Use case 9: Manipulate the available formats

- Use case 10: Manipulate the available encodings
- Use case 11: Manipulate the available languages

7.2 Work Assignments and Future Plans

The work plan for the next months are:

UNED will release the 1st Version of the MRM Toolkit with the functionalities showed above, working with NCSR in the integration of the toolkit within the Aqua system.

TKK will study the use of MMTx and MetaMap evaluating the features of the UMLS databases in order to decide the best way of designing the lexical resources repository. Also the possibility of extending these tools to other languages is required.

Finally, NCSR will help in integration issues and the design of the MRM databases.

Among our future plans, we are examining the possibility of considering the MRM repository as a global indexing platform for all the resources used in MedIEQ. Apart from the purely linguistic resources (vocabularies, lexicons, thesauri, ontologies), other resources could be handled by MRM such as the corpora to train/test the classification algorithms or even the configuration files needed by the different MedIEQ applications in all toolkits. This idea needs further analysis among all technical partners.

8. Concluding Remarks

This Deliverable D10 details the work done in the Methodology and Architecture for Multilingual Resources Management (WP7) during the first stage of the project, covering 2006.

We have compiled the list of linguistic resources and NLP tools owned by the MedIEQ partners and evaluated their feasibility for our purposes and needs.

We have proposed a set of linguistic and semantic resources and NLP tools over the health domain in order to help our expert users in their accreditation tasks of medical content. Since most of these resources and tools are available in the UMLS project, MRM toolkit's goal is to build an extra-layer on top of UMLS API and to be integrated in the Aqua system.

As some of the capabilities are still under development and testing stages, the presented architecture is still subject to change.

APPENDIX A: Further Details about Key Resources and Tools

Eurovoc

What is it?: Eurovoc is a multilingual thesaurus covering the fields in which the European Communities are active; it provides a means of indexing the documents in the documentation systems of the European institutions and of their users.

Users: The European Parliament, the Office for Official Publications of the European Communities, the national and regional parliaments in Europe, some national government departments and European organisations are currently using this controlled vocabulary.

Developers: The European institutions, the national parliaments and the various users of Eurovoc have cooperated to produce it.

Languages: Eurovoc 4.2 exists in 18 official languages of the European Union (Spanish, Czech, Danish, German, Greek, English, French, Italian, Latvian, Lithuanian, Hungarian, Dutch, Polish, Portuguese, Slovak, Slovene, Finnish and Swedish) and three another languages (Bulgarian, Romanian and Croatian).

ICD-10

What is it? ICD is a classification of diseases and other health problems recorded on many types of health and vital records including death certificates and hospital records. In addition to enabling the storage and retrieval of diagnostic information for clinical and epidemiological purposes, these records also provide the basis for the compilation of national mortality and morbidity statistics by WHO Member States.

The ICD has become the international standard diagnostic classification for all general epidemiological and many health management purposes. These include the analysis of the general health situation of population groups and monitoring of the incidence and prevalence of diseases and other health problems in relation to other variables such as the characteristics and circumstances of the individuals affected.

Users: ICD-10 was endorsed by the Forty-third World Health Assembly in May 1990 and came into use in WHO Member States as from 1994. The classification is the latest in a series which has its origins in the 1850s. The first edition, known as the International List of Causes of Death, was adopted by the International Statistical Institute in 1893.

Developers: WHO took over the responsibility for the ICD at its creation in 1948 when the Sixth Revision, which included causes of morbidity for the first time, was published.

Languages: ICD-10 exists in German, Dutch, and American English equivalent.

MeSH

What is it? Medical Subject Headings (MeSH) is a huge controlled vocabulary (or metadata system) for the purpose of indexing journal articles and books in the life sciences. MeSH has a strong clinical bent, making non-clinical searches, such as those being done to support epidemiological studies, more difficult than the norm. The 2005 version of MeSH contains a total of 22,568 subject headings, also known as descriptors. Most of these are accompanied by a short definition, links to related descriptors, and a list of synonyms or very similar terms (known as entry terms). Because of these synonym lists, MeSH can also be viewed as a thesaurus.

Users: it is used by the MEDLINE article database and by NLM's catalog of book holdings. MeSH can be browsed and downloaded free of charge on the Internet; a printed version is published once a year. The vocabulary and its supporting informatics systems were designed to be used both by indexing professionals and by medical staff with various degrees of computer experience. Using the vocabulary in support of database searches with the goal of scientific research often requires the help of specialized subject librarians.

Developers: Mesh was created and updated by the United States National Library of Medicine (NLM).

Languages: MeSH exists in English, Czech, Portuguese, Spanish, Finnish, German, Italian, Japanese, Dutch, Russian and Swedish.

MetaMap

What is it?: MetaMap is a highly configurable program that maps biomedical text to concepts in the UMLS Meta-thesaurus. This program includes the next steps:

- Parsing: Arbitrary text is parsed into (mainly) simple noun phrases.
- Variant Generation: For each phrase, variants are generated using the knowledge in the SPECIALIST lexicon and a supplementary database of synonyms.

- Candidate Retrieval: The candidate set of all Metathesaurus strings containing at least one of the variants is retrieved.
- Candidate Evaluation: Each Metathesaurus candidate is evaluated against the input text by first computing a mapping from the phrase words to the candidate's words and then calculating the strength of the mapping using a linguistically principled evaluation function consisting of a weighted average of four metrics.
- Mapping Construction: Complete mappings are constructed by combining candidates involved in disjoint parts of the phrase, and the strength of the complete mappings is computed just as for candidate mappings.

Developers: MetaMap has been developed by Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health and Department of Health and Human Services.

Languages: English and French

MMTx

What is it?: MMTx is an effort to make the MetaMap program available to biomedical researchers in a generic, configurable environment. MetaMap maps arbitrary text to concepts in the UMLS Metathesaurus; or, equivalently, it discovers Metathesaurus concepts in text.

With this software, text is processed through a series of modules. First it is parsed into components including sentences, paragraphs, phrases, lexical elements and tokens. Variants are generated from the resulting phrases. Candidate concepts from the UMLS Metathesaurus are retrieved and evaluated against the phrases. The best of the candidates are organized into a final mapping in such a way as to best cover the text.

Developers: MMTx has been developed by Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health and Department of Health and Human Services.

MTI

What is it? The MTI system consists of software for applying alternative methods of discovering MeSH headings for citation titles and abstracts and then combining them into an ordered list of recommended indexing terms. The top portion of the diagram consists of three paths, or methods, for creating a list of recommended indexing terms: MetaMap Indexing, Trigrams and PubMed Related Citations. The first two paths actually

compute UMLS Metathesaurus concepts which are passed to the Restrict to MeSH process. The results from each path are weighted and combined using the Clustering process. The system is highly parameterized not only by path weights but also by several parameters specific to the Restrict to MeSH and Clustering processes.

Developers: MTI has been developed by Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health and Department of Health and Human Services.

Languages: English

SNOMED

What is it? The Systematized Nomenclature of Medicine is a system of standardized medical terminology. SNOMED is a comprehensive and precise clinical reference terminology that provides unsurpassed clinical content and expressivity for clinical documentation and reporting. It allows a consistent way to index, store, retrieve, and aggregate clinical data across specialties and sites of care. It also helps structure and computerize the medical record, reducing the variability in the way data is captured, encoded and used for clinical care of patients and research. SNOMED created a common clinical language that is a necessary element of a health care Information Infrastructure. SNOMED-CT cross maps to such other terminologies as ICD-9-CM, ICD-O3, ICD-10, Laboratory LOINC and OPCS-4. It supports ANSI, DICOM, HL7, and ISO standards.

Developers: SNOMED was developed by the College of American pathologists (CAP).

Users: Clinicians and organizations in order to solve the problem of using different clinical terms that mean the same thing.

Languages: English, German and Spanish.

UMLS

What is it? The Unified Medical Language System (UMLS)⁹ is a set of controlled vocabularies which also provides a mapping structure between them, allowing to create, process, retrieve, integrate, and/or aggregate biomedical and health data and information. The National Library of Medicine (NLM) produces and distributes the UMLS Knowledge Sources (databases) and associated software tools (programs)

The UMLS is composed of three main components:

⁹ Unified Medical Language System (UMLS)'s homepage:
<http://umlsinfo.nlm.nih.gov/>

- The *Metathesaurus* is the base of the UMLS and it is comprised of over 1 million biomedical concepts and 5 million concept names, all of which are from over 100 controlled vocabularies and classification systems used in patient records, bibliographic, administrative health data and full text databases. Some examples of the controlled vocabularies are ICD, MeSH, SNOMED CT, LOINC, and RxNORM.

The *Metathesaurus* is organized by concept or meaning, and each concept has specific attributes that define the meaning. Identical or almost identical concepts are linked together with hierarchical context from the different vocabularies and relationships between the concepts are explained and represented.

- The *Semantic Network*¹⁰ has Semantic types and Semantic relationships, that exist between semantic types. There are 135 semantic types and 54 relationships. This network is designed to categorize concepts in the UMLS Metathesaurus and provide relationships among the concepts. Once a Metathesaurus concept is established, it is connected to the most specific semantic type from the Semantic Network.
- The SPECIALIST Lexicon is the third of the Knowledge Sources supporting the Unified Medical Language System. Both common English vocabulary and biomedical terms are a source for the Specialist Natural Language Processing System, as well as information from MEDLINE¹¹, and the UMLS Metathesaurus. Each entry contains syntactic (how words are put together to create meaning), morphological (form and structure) and spelling information.

UMLSKS

What is it? UMLS Knowledge Source Server (UMLSKS)¹² is a computer application that provides Internet access to the Knowledge Sources and other related resources made available by developers using the UMLS.

Users: technical partners involved in WP7.

10A semantic network is a knowledge representation schemes involving nodes and links (arcs or arrows) between nodes. The nodes represent objects or concepts and the links represent relations between nodes. This graphical representation assists in understanding the relationships of concepts. For further details, please see http://en.wikipedia.org/wiki/Semantic_networks

11 Medical Literature Analysis and Retrieval System Online (MEDLINE)'s homepage: <http://www.ncbi.nlm.nih.gov/entrez>

12 UMLS Knowledge Source Server (UMLSKS)'s homepage: <http://umlsks.nlm.nih.gov/kss>

UMLS^{SKS}'s purpose is to make UMLS data more accessible to users, and in particular to system developers. The system architecture allows remote site users (individuals as well as computer programs) to send requests to a server at the National Library of Medicine (NLM). Access to the system is provided through the World Wide Web, an Extensible Markup Language (XML)-based socket programming interface, and through an Application Programmer Interface (API).

The new Application Programmer Interface (API) is entirely written in Java with the goal of providing a platform independent form. In addition, an XML-based API resembling the Java API methods is included allowing both Java and non-Java programs to interface to the UMLS through a standard TCP socket. The flexibility provided by both APIs enables the UMLS^{SKS} to support developers whose platforms range from PCs to high-end Unix machines. In all, approximately 40 API methods have been defined allowing access to all details of the Metathesaurus.

The API download¹³ includes Javadoc documentation for all interface and object model classes¹⁴, a set of example Java programs for issuing API calls, some sample XML documents that may be used as input to the UMLS^{SKS} socket interface, and sample XML output files for each of the API methods.

13. UMLS^{SK} API download:

<http://umlsks.nlm.nih.gov/kss/servlet/Turbine/template/docs%2Capi%2CapiDownload.vm>

14 UMLS^{SKS} API's Javadoc: <http://umlsks.nlm.nih.gov/kss/api/html/index.html>