

SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

D20.3

Platform for Exchange of Biomedical Text Corpora

Report Version: 0.1

Report Preparation Date: January 27, 2006

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	THE CORPUS REPOSITORY EXPLORER	4
4.1	Motivation for the Need of an Exchange Platform.....	4
4.2	Corpus Acquisition.....	4
4.3	Basic Structure	5
4.4	Uploading A Corpus.....	6
4.5	Browsing the Corpus.....	8
4.6	Integration into DocMan and Adaptation.....	9
4.7	Final Release and Future Work	10
5	CONCLUSION	10

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:	D20.3
Authors:	Ekaterina Buyko (JENA), Michael Poprat (JENA), Stefan Schulz (UKLFR)

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 an 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.

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, corpus-based machine translation and corpus management. 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.
- To provide a Web-based platform for automated processing for corpora and a up- and downloading interface for project partners (restricted use)

4 The Corpus Repository Explorer

4.1 Motivation for the Need of an Exchange Platform

During the last years, the members of WP 20 have been collecting a huge amount of different corpora of different formats from different domains (e.g., biomedical texts, clinical texts, newspaper texts etc.) and written in different languages (German, English, Spanish, Swedish etc.). These texts have been used for different studies in the area of medical informatics and NLP applications, like building multilingual dictionaries, term extraction, training corpora for unsupervised learning methods etc.

Unfortunately, this document collection has been stored at different locations in several file systems without following a consistent classification scheme. However, this would be an important help to avoid unnecessary acquisition efforts (writing tools for downloading documents, cleaning them etc.) and to avoid redundancy. These aspects are even more crucial when corpora should be made available for other groups (e.g. the member of the NoE “Semantic Mining”). Here, the access to the file system is restricted. Repository systems like CVS or Subversion are too sophisticated. As a consequence, a shared repository of corpora should be centralised, provide an easy (but still controlled) access and straightforward up- and downloading facilities.

In a first try, UKLFR and JENA prepared guidelines for corpus acquisition, management and documentation. In this draft, we focussed on problems and hints for the manual download of corpora (see also section *Corpus Acquisition*), we elaborated a repository hierarchy (see section *Basic Structure of the Corpus Hierarchy* and we provided an XML structure for meta data which describe each corpus. Furthermore, we proposed different processing steps and versions of different processing levels a corpus should provide. Although we clearly described the handling with corpora, we had problems to convince the users to follow these guidelines, e.g. to strictly separate the repository from tools and individually generated data. It proved impossible to enforce these guidelines in a normal file system. As a consequence, the planned strict repository structure could not be realised.

In September 2005, we decided to revise the guidelines in order to reduce the amount of manual work. The idea was that once a corpus was mirrored from the Web, the subsequent processing steps as well as the organisation of the documents in the proposed hierarchy can be automated. The access to this repository should only be possible via a Web interface, both for uploading and downloading. The direct user access to the directories is no longer possible.

Up to now, a prototype of the corpus repository system (the Corpus Explorer) was implemented. In the existing version, we emphasised functionality and clearly disregarded handling and user interface design. We plan the integration of the corpus repository software into an existing document management system in order not to care about the points mentioned before.

4.2 Corpus Acquisition

The acquisition of documents from the Web is a bold venture for at least two reasons. First, some owners of websites forbid mirroring their pages on a local file system, others restrict the use for private purposes only and some do not give any explicit hint. We here often move in a semi-legal area. As a rule of thumb, we assume the download of all pages to be legal that can also be reached by a search robot. Pages that are protected by username and password are not allowed to download

without having asked the owner for permission. Second, the acquisition of pages from the Web may generate a lot of data traffic resulting in high bills or in crashes of Web servers. To avoid this problem, the selection of pages to be downloaded must be done carefully. This is the main reason why an automatic download of pages cannot be realised. In addition, a manual examination of the organisation of a Web page is required when we want to access only particular sites (e.g., print versions of pages). We here propose some tools and approaches for simply downloading the desired pages. Mostly, these tools have to be integrated in shell scripts, Perl programmes etc.

A very useful tool, generally available on every Linux or Unix installation, is *wget* (for Windows users, this tool can be downloaded separately or is part of the Linux simulation cygwin). In principal, *wget* starts with a certain page and follows every link until a depth of n is reached. Here, we should act with caution: As a link will be followed recursively, the amount of pages to be downloaded can get out of control, especially when following external links. It is very recommendable to set constraints (like *wget -mirror URL*). Furthermore, every non-textual element (like pictures) should be excluded to decrease the amount of traffic. For more detailed information, we refer to the corresponding documentation in the man-pages. In any case, one should assure that the storage system where the sites are downloaded provides enough memory (for example by *df -ksh .*).

4.3 Basic Structure

The basic structure of the corpus repository is organised in three layers (see Figure below). First, the top level in this hierarchy distinguishes between the **genre** of the corpus (biomedical vs. newspaper). In the second level, the documents are categorised according to their **language**. At them moment, we have biomedical corpora in German, English, French, Portuguese, Spanish and Swedish. Below this language level, the corpora are organised by their **identifier** (as a rule, the identifier correspond to their Web source like *www.netdoctor.co.uk* or *www.mayoclinic.com*). Beyond this level, the creation of directories will be performed automatically. Note that, at the moment, the kind of organisation is also reflected by the level of directories. For the future work, the levels will be realised on a more abstract level, depending on the document management tool we envisage to use.

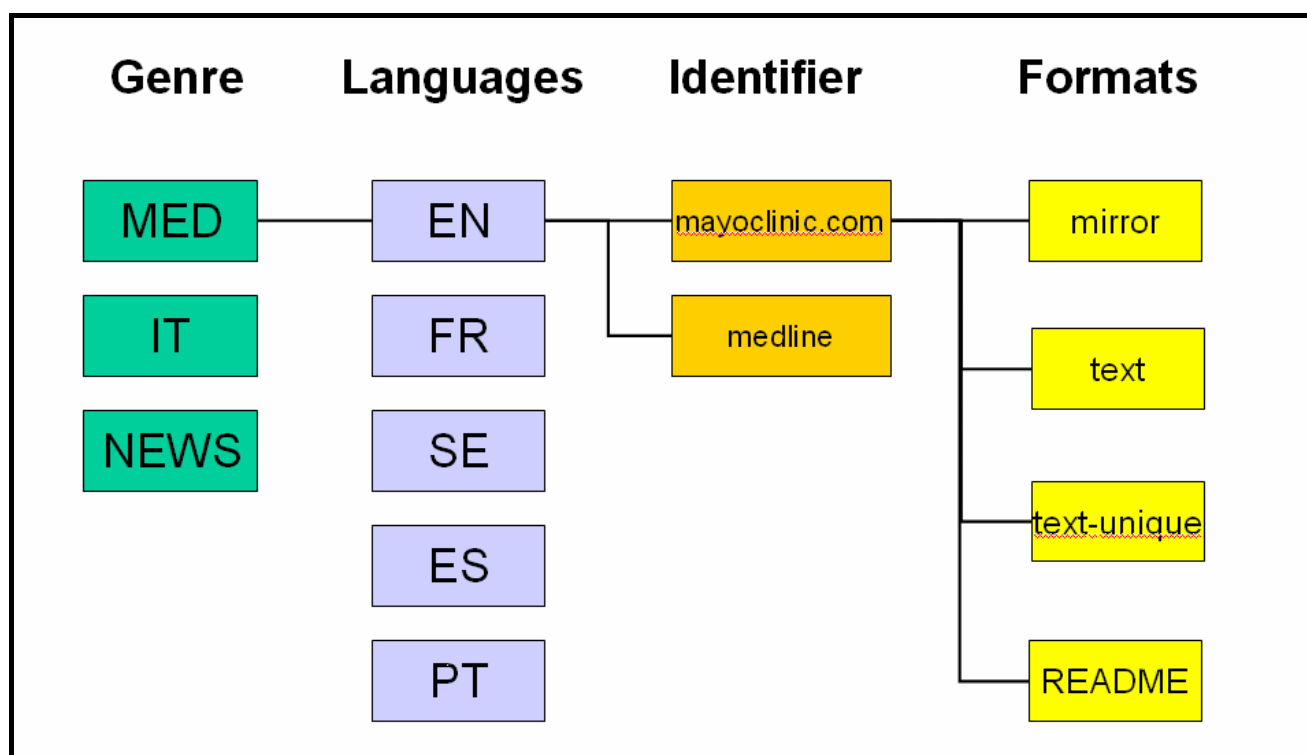


Figure 1: Basic Structure of the Copus Repository

4.4 Uploading A Corpus

Having downloaded a corpus as described in chapter *Corpus Acquisition*, we first have to choose the appropriate category where the corpus should be categorised (see Figure X). At the moment, the corpus has to be copied in the mirror-directory manually (a point that will be obsolete as soon as the corpus repository will be integrated in the document management software, see also chapter *Integration into DocMan*). Some meta-information about source, date, restriction level and comments can be provided in a form presented by the corpus tool. At the same time, the automatic processing of the corpus and the creation of the different corpus formats will be started, performing the following steps:

- **Elimination of meta tags (html, xml etc.), extraction of the relevant text and converting to UTF-8:**
Instead of simply using `lynx --dump`, which allows us to extract the text without html tags (this was proposed in the first guidelines), we have elaborated a more sophisticated tool. Now, we are able to identify menu structures heuristically and to ignore them for the textual output. Furthermore, if the codepage encoding is given in the header of an html page, we are able to convert the text into UTF-8. We also try to extract long lines (only broken by line breaks or paragraphs), which is important for the elimination of redundant lines (see next point).
- **Elimination of redundant lines:**
In order not to bias the textual information with redundant lines (like menu points to recognized with the previous tool, copyright statements etc.), we provide a corpus version containing only unique lines. We here use the UNIX tool `sort -u`, piping all documents into a large one. This step is essential for all applications that rely on the occurrence and distribution of n-grams of tokens.
- **Tokenisation of the corpus (not realised yet) :**
It had been previously proposed to provide all corpora in a linguistically pre-processed form

(tokenized, chunked, parsed). This is disadvantage for many reasons, like non-standardized tag sets, non-existing models for every domain etc. Even for the apparently simple task of tokenization, it is nearly impossible to define a standard. Even so, in the near future, we will offer all corpora in a basic (and very radical) tokenised form that is independent of any task and domain.

For every processing step, a zipped version of the corpus will be created and made available for download.

Most of the programmes just described are implemented in the programming language *Python*, while the corpus repository itself is written in *PHP*. *PHP* is also the language of the content and document management tool where the current corpus repository tool will be integrated.

The screenshot shows a web browser window titled "Add Corpus - Mozilla Firefox". The address bar contains the URL "http://supreme.coling.uni-jena.de/corpora/addRepository.php". The main content area has the heading "Add Corpus". Below the heading, there is a section "Source of mirror directory (Examples):" with a text input field and three radio button options: "/data/data/data_corpora/prototype/english/medline_small/", "/data/data/data_corpora/prototype/german/msd_doc/", and "/data/data/data_corpora/prototype/swedish/www.fass.se/". The next section is "README.XML" with fields for "Source:" (containing "PubMed"), "Date:" (containing "17.01.2006"), and "Who:" (containing "Ekaterina Buyko"). Below that is "Level of Restriction:" with three radio button options: "not legal", "not demanded", and "legal" (which is selected). The final section is "Comments (max. 300 characters):" with a large text input field.

Figure 2: Corpus ExplorerUpload Form (Prototype)

4.5 Browsing the Corpus

Browsing the corpus repository is very simple: to get a particular corpus, one just has to follow the repository structure described before. We start with choosing the desired genre, then the language and finally one of the offered corpora. Here, we get information about the corpus (that has been stored in the READMED.XML) and some automatically generated statistics like size of the (zipped) corpus, number of documents and number of words (based on the original corpus). At this level, the user has the possibility to download a zipped version.

Repository Browser - Mozilla Firefox

Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe

http://supreme.coling.uni-jena.de/corpora/browseRepository.php?dir=/data/data/data_corpora/prototype//english

Erste Schritte Aktuelle Nachrichten...

Repository Browser

Path: /data/data/data_corpora/prototype//english/medline_small

[UP](#)

Corpus Info:	
Source:	PubMed
Date:	17.01.2006
Who:	Ekaterina Buyko
Restriction Level:	<input type="radio"/> not legal <input type="radio"/> not asked <input checked="" type="radio"/> legal
Comments:	
Number of Documents:	55
Number of Tokens:	25292

[mirror.zip \(319.82 KB\)](#)
[mirror text.zip \(79.58 KB\)](#)
[text unique.zip \(38.18 KB\)](#)

Figure 3: Corpus Explorer Browsing and Download Centre

4.6 Integration into DocMan and Adaptation

For the Web presentation of the JENA group, we decided to rely on a content management system called *Joomla* (former known as *Mambo*). Additional existing tools for different tasks, the so called modules, can be integrated very easily into this system. For the organisation of the documents which should be available for the public (like slides and exercise sheets for lectures), we use a document management tool called [DocMan](#). This tools allows (registered) users to download and upload documents into a given categories. The documents can be provided with comments as well as automatically generated meta information like document size, creator, date etc.

The screenshot shows the Joomla CMS interface with the DocMan module. The main content area displays a list of documents under the heading 'Downloads'. The selected document is 'Übungsblatt 5 - Lösungsvorschlag', and its details are shown in a tooltip table.

Property	Value
Name	Übungsblatt 5 - Lösungsvorschlag
Description	
Filename	loesung_blat05.doc
Filesize	37,5 KB
Filetype	doc (Mime Typeapplication/msword)
Creator	admin
Created On:	20.12.2005 13:01
Viewers	Everybody
Maintained by	
Hits	15Hits
Last updated on	20.12.2005 13:02
Homepage	

Figure 4: Application of DocMan as a module in the Joomla CMS

The clear advantages of DocMan lie in the already existing, clearly arranged graphical interface. Furthermore, the possibility of nesting categories without a direct access to the file system makes this solution more secure than the current repository software. Also the possibility to upload documents from the local (and remote!) file system is a desirable feature. And, also very important, up- and download of particular documents can be restricted to particular users. This would allow us to make some of the corpora available to the public, others to be used only internally and, in turn, some of them only for group members.

Although most basic features are available, we need to adapt the DocMan software to some of our purposes. Most importantly, the corpora processing scripts for the creation of the different versions

(see chapter *Adding a Corpus*) have to be integrated in the upload-feature of DocMan. In addition, DocMan in its current version allows uploading only one file at a time. We here have to realise a recursive upload of complete directories, the pages we mirrored as described in chapter *Corpus Acquisition*. Another feature to be implemented is the creation of categories via the DocMan front-end according to the given basic hierarchical structure. This is always the case when uploading a new corpus, but also when we want to include a new genre or a new language. Moreover, the information coming with the corpus must be adapted.

4.7 Final Release and Future Work

Beside the adaptation of DocMan, some other corpus repository features are pending. E.g., some corpus repositories also contain PDF files, which should also be processed and converted like html files. Moreover, at the moment, the starting point of all processing steps has to be a repository with html (incl. xml etc.) files. Here, an automatic detection of the file type and an adequate processing would be helpful. Concerning the UTF-8 conversion problems, our tool fails if the character encoding is unspecified in the source. We plan to integrate an enhanced programme that guesses the codepage of a text document.

In the future, we want to provide a version for each corpus in a tokenised form. EBI and JENA are currently working on rules and a programme for a basic tokenisation. This will be a further step in the processing pipeline.

Finally, certain corpora still do not fit into our categorisation scheme, e.g. manually annotated corpora like treebanks. We will probably create a new branch at the top level for this kind of documents.

5 Conclusion

The corpus repository project was presented at the work package meetings in Paris (Sept. 2005) and Geneva (Jan. 2006). The feedback was positive and the tools seem to be of interest for most partners. Feedback from the audience was compiled into the tool development and in the to-do list.

Furthermore, the new work package WP 27 (making patient records understandable for laymen) mentions the Corpus Repository as an essential requirement for their work. We here intend further collaborations with WP27.

Nevertheless, the current status of the Corpus Explorer is still prototypical. So it can be used only for internal purposes. The tool will be functional in February 2006, after cross-checking with the requirements of the new workpackage 27.

Corpus Explorer will be made available on the Web at <http://supreme.coling.uni-jena.de/corpora> (username: guest; password: julie2006)